
# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**UNDERSTANDING THE IMPACT OF SOCIO-ECONOMIC FACTORS ON NAVY ACCESSIONS**

by

Bradley C. Intrater

September 2015

| | |
|---|---|
| Thesis Advisor: | Jonathan K. Alt |
| Second Reader: | Samuel E. Buttrey |

THIS PAGE INTENTIONALLY LEFT BLANK

# REPORT DOCUMENTATION PAGE

*Form Approved OMB No. 0704–0188*



| **1. AGENCY USE ONLY** ***(Leave blank)*** | **2. REPORT DATE** September 2015 | **3. REPORT TYPE AND DATES COVERED** Master's Thesis |
|---|---|---|

**4. TITLE AND SUBTITLE**
UNDERSTANDING THE IMPACT OF SOCIO-ECONOMIC FACTORS ON NAVY ACCESSIONS

**5. FUNDING NUMBERS**

**6. AUTHOR(S)** Intrater, Bradley C.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Naval Postgraduate School
Monterey, CA 93943-5000

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Navy Recruiting Command (NRC)
5722 Integrity St #784
Millington, TN 38054

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES** The views expressed in this thesis were those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. government. IRB Protocol number ____N/A____.

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT (maximum 200 words)**

In a fiscally constrained environment, Navy Recruiting Command (NRC) must assign its recruiters to maximize the annual number of accessions by each recruiting station. Our thesis built on research in this area and made use of open source socio-economic data from several sources, including the Internal Revenue Service (IRS) and the Federal Bureau of Investigation (FBI). Beginning with a response variable of annual Navy accessions and a set of 71 independent predictor variables populated from ZIP code-level data, we fit and validated six predictive regression models. Models were fit using multiple linear regression (MLR) at the station level and zero-inflated negative binomial (ZINB) regression at the ZIP code level. We identified average number of recruiters, adjusted gross income (AGI) < $25,000, and total veterans as the principal drivers of accession production. We identified AGI > $200,000, unemployment compensation, and total number of universities in a ZIP code as the principal inhibitors to accessions. With out-of-sample data and using 95% prediction intervals, we tested the performance for each of the MLR models and validated them using the five assumptions of linear models. We tested the ZINB models against an out-of-sample subset using Mean Absolute Deviation (MAD) and true negatives, which verify the prediction rate of structural and random zeros. MAD and true negatives demonstrated improvement from previous zero-inflated Poisson models developed in 2011 by Y. K. Pinelis, E. Schmitz, Z. Miller and E. Rebhan, of the Center for Naval Analysis (CNA), in *An Analysis of Navy Recruiting Goal Allocation Models.*

| **14. SUBJECT TERMS** Multiple-linear regression, zero-inflated Poisson, zero-inflated negative binomial, Logistic Regression, Navy Recruiting Command, variable selection, socio-economic data | **15. NUMBER OF PAGES** 147 |
|---|---|
| | **16. PRICE CODE** |

| **17. SECURITY CLASSIFICATION OF REPORT** | **18. SECURITY CLASSIFICATION OF THIS PAGE** | **19. SECURITY CLASSIFICATION OF ABSTRACT** | **20. LIMITATION OF ABSTRACT** |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UU |

NSN 7540–01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

# UNDERSTANDING THE IMPACT OF SOCIO-ECONOMIC FACTORS ON NAVY ACCESSIONS

Bradley C. Intrater
Lieutenant, United States Navy
B.A., University of Colorado, Boulder, 2007

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the

**NAVAL POSTGRADUATE SCHOOL**
**September 2015**

Author: Bradley C. Intrater

Approved by: LTC Jonathan K. Alt
Thesis Advisor

Dr. Samuel E. Buttrey
Second Reader

Dr. Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

In a fiscally constrained environment, Navy Recruiting Command (NRC) must assign its recruiters to maximize the annual number of accessions by each recruiting station. Our thesis built on research in this area and made use of open source socio-economic data from several sources, including the Internal Revenue Service (IRS) and the Federal Bureau of Investigation (FBI). Beginning with a response variable of annual Navy accessions and a set of 71 independent predictor variables populated from ZIP code-level data, we fit and validated six predictive regression models. Models were fit using multiple linear regression (MLR) at the station level and zero-inflated negative binomial (ZINB) regression at the ZIP code level. We identified average number of recruiters, adjusted gross income (AGI) < $25,000, and total veterans as the principal drivers of accession production. We identified AGI > $200,000, unemployment compensation, and total number of universities in a ZIP code as the principal inhibitors to accessions. With out-of-sample data and using 95% prediction intervals, we tested the performance for each of the MLR models and validated them using the five assumptions of linear models. We tested the ZINB models against an out-of-sample subset using Mean Absolute Deviation (MAD) and true negatives, which verify the prediction rate of structural and random zeros. MAD and true negatives demonstrated improvement from previous zero-inflated Poisson models developed in 2011 by Y. K. Pinelis, E. Schmitz, Z. Miller and E. Rebhan, of the Center for Naval Analysis (CNA), in *An Analysis of Navy Recruiting Goal Allocation Models.*

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

# LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| AC | active component |
| ACS | American Community Survey |
| AGI | adjusted gross income |
| AIC | Akaike Information Criterion |
| AOR | area of responsibility |
| ASAD | all service accession data |
| ASCD | all service contract data |
| ASVAB | Armed Services Vocational Aptitude Battery |
| BIC | Bayesian Information Criterion |
| CAN | Center for Naval Analysis |
| CNRC | Commander Navy Recruiting Command |
| DMDC | Defense Manpower Data Center |
| DOD | Department of Defense |
| EGM | enlisted goaling model |
| FBI | Federal Bureau of Investigation |
| FIPS | Federal Information Processing Standard |
| GI | Government Issue |
| GLM | Generalized Linear Model |
| HBCU | historically black college or university |
| HSDG | high school diploma graduate |
| IPEDS | Integrated Postsecondary Education Data System |
| IRS | Internal Revenue Service |
| IV | independent variable |
| JAMRS | Joint Advertising Market Research and Studies |
| MAD | mean absolute deviation |
| MEPS | Military Entrance Processing Station |
| MLR | multiple-linear regression model |
| NA | not applicable |

| | |
|---|---|
| NALTS | National Advertising Leads Tracking System |
| NAVCRUITCOM | Navy Recruiting Command |
| NLL | negative log likelihood |
| NRC | Navy Recruiting Command |
| NRD | Navy Recruiting District |
| NRR | Navy Recruiting Region |
| NRS | Navy Recruiting Station |
| OLS | ordinary least squares |
| PI | prediction interval |
| PMF | probability mass function |
| PRIDE | personalized recruiting for immediate and delayed enlistment |
| PSR | Personnel Status Report |
| PYTHON | general-purpose programming language |
| QMA | qualified military available |
| R | statistical programming language |
| RAF | recruiter assignment factor |
| RFMIS | Recruiting Facilities Management Information System |
| RSID | Recruiting Station Identification Number |
| RSS | residual sum of squares |
| SAMA | Segmentation Analysis and Market Assessment |
| SMART | station market analysis and review techniques |
| SOI | statistics of income |
| STD | student testing program data |
| STEAM | standardized territory evaluation and analysis for management |
| TSC | test score category |
| TSS | total sum of squares |
| USAREC | United States Army Recruiting Command |
| USMEPCOM | United States Military Entrance Processing Command |
| VET | veteran |
| VIF | variance inflation factors |
| W&P | Woods and Poole |
| WWII | World War Two |

| | |
|---|---|
| YATS | Youth Attitude Tracking Studies |
| ZINB | zero-inflated negative binomial |
| ZIPR | zero-inflated Poisson regression |
| ZIP | Zone Improvement Plan |

THIS PAGE INTENTIONALLY LEFT BLANK

# EXECUTIVE SUMMARY

Navy Recruiting Command (NRC) recruits in all U.S. states and territories. The Command's hierarchy consists of 952 (as of May 2015) Recruiting Station IDs (RSIDs), which comprise the 26 Navy Recruiting Districts (NRD). These NRDs are divided into two regions, east and west, and those two regions together cover NRC's area of responsibility (AOR). With the vastness of NRCs AOR and its 30,000+ ZIP codes comes difficulty in understanding the complexity of the constantly changing demands and knowing precisely where to assign recruiters in order to maximize penetration into the market.

With the proliferation of governmentally collected socio-economic data, the potential exists to use statistical modeling and analysis to provide NRC leadership greater insights into an area's recruiting potential and the important variables impacting accessions. Station and ZIP code-level models will aid both, district leadership and station leadership, to utilize their limited resources efficiently.

All original data sets were collected from open source websites, including IPEDS, IRS, FBI, and the U.S. Census Bureau. Data were selected based on availability, accessibility, quality, and the previous research in this area. Three years of data were collected. We considered two levels of analysis for this research: ZIP codes and stations. Since stations consist of a collection of ZIP codes, with each ZIP code being assigned to only one station, a methodology was required to aggregate the data to the station level. Depending on the data, the mean, median, maximum, or sum across all ZIP codes assigned to the station was used at the station level. A total of 71 independent variables were prepared to support model fitting.

Following compilation of all data, two subsets were created. An in-sample data subset from the 2011 data was created from which to fit the models. Another out-of-sample data subset from the 2012 data was created as the test set and used in the performance test phase.

Based on previous research and an initial analysis of the predictor variable trends, we used multiple linear regression (MLR) and zero-inflated negative binomial (ZINB) regression to train models to predict accessions and determine the most statistically significant independent variables (Marmion, 2015; Zeileis, 2008). MLR modeling technique describes a linear relationship between a dependent response ($y_i$) variable and multiple independent explanatory variables ($x_{in}$) using ordinary least squares (OLS) estimates (Faraway, 2006). The ZINB model combines two components jointly into one simultaneous model using a binomial and count model (Chin, 2015).

The model development process consisted of multiple systematic reviews using a series of statistical tests and criteria to select the best-fitting subset of model variables and variable transformations to improve the goodness-of-fit while minimizing multi-collinearity. Independent variables (IV) with a p-value < .01 were considered statistically significant.

We fit MLR station level models using the dropterm() function in R (Venables & Ripley, 2002). Using minimum Bayesian Information Criterion (BIC) as a variable discriminator, this function allows modelers to systematically drop one-term-at-a-time to select and discard appropriate variables for the final fitted model (R Core Team, 2013). We fit ZINB ZIP code-level models using the be.zeroinfl() function in R (Zhu Wang, 2015). Using a significance level of p-value < .01 as a variable discriminator, this function selects and eliminates variables until the best fitting model is achieved.

Following variable selection and model diagnostics, models were tested against out-of-sample data sets. Results from the performance tests of the MLR and ZINB models are in Tables 1 and 2.

Table 1. Summary statistics for goodness-of-fit (adjusted R2) of each regional model (national, east, and west), counts for the observations captured by the 95% prediction interval (PI), the fraction of observations in that PI, and the number of independent variables in the model. For number of observations within the 95% PI, we divided the observations within the PI by the total number of observations in that region.

| | In-Sample Adjusted $R^2$ | Out-of-Sample Adjusted $R^2$ | Number of observations ($y_i$) within 95% PI | Percent of observations ($y_i$) within 95% PI | Number of independent variables in model |
|---|---|---|---|---|---|
| National | 59% | 46% | 862/952 | 91% | 7 |
| East | 60% | 43% | 425/465 | 91% | 6 |
| West | 56% | 43% | 446/487 | 91% | 5 |

Table 2. Summary view of the metrics used in testing the performance of the ZINB models.

| Metric | National ZINB | East ZINB | West ZINB |
|---|---|---|---|
| Model MAD | 0.73 | 0.732 | 0.692 |
| Count MAD | 1.65 | 1.47 | 1.78 |
| True Positive | 0.279 | 0.297 | 0.254 |
| True Negative | 0.808 | 0.761 | 0.859 |
| False Positive | 0.192 | 0.239 | 0.141 |
| False Negative | 0.177 | 0.165 | 0.192 |
| Number of predicted zeros | 8899 | 9045 | 9805 |
| % of zeros | 0.514 | 0.522 | 0.626 |

Model Mean absolute deviation (MAD) is a measure of the average accuracy for the overall ZIP code model. Count MAD measures the average accuracy of the model without structural and random zeros. A smaller MAD is desired. Positive is defined as an observation with at least one accession. True is defined when the model's prediction matches the actual observation.

Insights and recommendations are as follows:

- Recruiter strength was the best predictor of accessions in the station model; however, the relationship between recruiters and accessions is not linear. The positive impact on accessions beyond 6 recruiters per station appears to reach a plateau.
- Wealthy (AGI > $200,000) areas with a large concentration of Division I universities produce fewer recruits.

- Recruiter strength, total veterans, underprivileged (AGI < $25,000) areas, and violent crime are the strongest positive predictors in the ZIP code models.

# LIST OF REFERENCES


Chin, D.-G. (2015). *Innovative statistical methods for public health.* New York: Springer.

Commander Navy Recruiting Command (CNRC). (2011, May 17). *Navy recruiting manual,recruiting operatons*. Millington TN: CNRC. Retrieved June 3, 2015, from COMNAVCRUITCOMINST 1130.8J: http://www.cnrc.navy.mil/Publications/Directives/1130.8/1130.8J_VOL%20I_Recruiting%20Operations-CH8.pdf

Faraway, J. (2006). *Extending the linear model with R*. Boca Raton, FL: Taylor and Francis Group.

Marmion, W. (2015). *Evaluating and improving the SAMA (Segmentation Analysis and Market Assessment) recruiting model* (Master's thesis). Retrieved from Calhoun http://calhoun.nps.edu/bitstream/handle/10945/45894.

Pinelis, Y. K., Schmitz, E., Miller, Z. & Rebhan, E. (2011). *An analysis of Navy recruiting goal allocation model*. Arlington, VA: Center For Naval Analysis.

R Development Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: The R Foundation for Statistical Computing

Rodriguez, G. (2013, November 6). Models for count data with overdispersion. Retrieved from Princeton University http://data.princeton.edu/wws509/notes/c4a.pdf

Venables, W. N. & Ripley, B. D. (2002). Modern applied statistics with S. Fourth Edition. New York: Springer.

Zeileis, A. (2008). Regression models for count data in R. *Journal of Statistical Software, 27* (8), 7.

Zhu Wang, with contributions from Achim Zeileis, Simon Jackman, Brian Ripley, Trevor Hastie, Rob Tibshirani, Balasubramanian Nrasimhan, Gil Chu and Patrick Breheny (2015). mpath: Regularized linear models. R package version 0.1-19.

# ACKNOWLEDGMENTS

Thank you to my parents, Rob and Dawn. I could not have completed this program without your guidance and encouragement. Regardless of my successes and failures as a child, you have always taught me to stay well rounded, and it has certainly helped me at Naval Postgraduate School. To my advisor, second reader, and professors, thank you for providing your expertise and helping me to keep my head to the grindstone. Most importantly, thank you to my wife, Elizabeth; you, Virginia, and Will have been a steady rock for me over the past two years, and your unwavering support through the long nights and early mornings have kept me going strong. This would not have been possible without you.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

## A. INTRODUCTION

In order to satisfy congressionally mandated end-strength goals, Navy enlisted recruiters must sift through the pool of qualified and unqualified military-age candidates to find the 29% of the 34 million youths, ages 17–24, in the United States who are eligible for military service (Feeney, 2014). Navy Recruiting Command (NRC) must operate within the boundaries of social, political, and economic constraints while maximizing use of its limited resources, namely money and human capital, to achieve this goal. The challenges in meeting this goal lie in identifying the factors, socio-economic and other, that influence an area's ability to generate Navy recruits so that Navy recruiters may position themselves in a manner that increases their expected recruit production.

The Selective Training and Service Act of 1940 ended in 1973 and the military became an all-volunteer force. Now, the Navy must compete with civilian industries, corporations, and universities for its nation's youth. Following the termination of the draft, many organizations and academic institutions conducted studies to identify the best recruiting markets within the United States (Gibson, 2009). With the proliferation of open-source socio-economic data sets from government websites and recognizing that the factors effecting the recruiting market change over time, NRC identified the need for updated statistical models of recruit production and a better understanding of potential insights into recruit production that might be gained from these data.

The principal objective of this research is to identify the socio-economic factors that affect Navy active component (AC) accessions, explore the development of statistical models that consider these factors, and characterize their performance compared to existing models. This research provides NRC's N5 (Research, Plans, and Analysis) Branch models at the Navy Recruiting Station (NRS) level and ZIP code-level using open-source ZIP code-level independent variable (IV) data. Effective use of these models could result in a better understanding of each district's recruiting market and a

potential cost savings to the Navy by improved alignment of recruiting stations and ZIP codes.

## B. PROBLEM STATEMENT

With more youth considering college and the general upward trend of the economy, the Navy will be confronted with the challenge of recruiting from a population of youths who have broadened opportunities (Cook, 2014). As these opportunities for youth become more numerous, the Navy will have difficulty recruiting an all-volunteer force (AVF) with continued manpower and fiscal cutbacks (D. Ammons-Moreno, personal communication, August 20, 2015). This research will develop statistical models that aid NRC in estimating the number of accessions a geographic area is expected to produce on an annual basis. If recruiters are not efficiently utilized and placed in the most prolific areas, NRC increases the risk of not making its set goals. Potential cost-cutting measures include closing low-yield stations and replacing them with virtual stations; but the conundrum lies in knowing precisely which ones to close. Accurately identifying non-productive ZIP codes is a problem that still requires a solution (Pinelis, Schmitz, Miller, & Rebhan, 2011).

In a fiscally constrained environment, recruiter placement and efficiency are vital to maximizing resources. Previously, the Navy and military as a whole relied on drafts and mass-marketing campaigns with an enormous pool of recruiters, some of whom would often "drive 1 to 1.5 hours to meet with potential recruits" (Pinelis et al., 2011, p. 8). This is a sizeable waste of resources, and the Navy, in this fiscally constrained environment, would be imprudent to continue this approach. NRC requires these statistical models to identify ideal recruiter allocation, station alignment organization, low-producing stations, and non-producing stations, and the socio-economic factors that affect accessions.

## C. RESEARCH OVERVIEW

This study is divided into five chapters. Chapter II covers the background of the requirement for recruiting models and those used by NRC and U.S. Army Recruiting Command (USAREC). It concludes with a literature review of related work. Chapter III

covers the methodology for collecting and processing the data and constraints, assumptions, and limitations for the models. Because we collected and cleaned all data from open sources, we will outline techniques used and methods for creating the final data sets. Chapter IV focuses on the model development, output, and analysis. Chapter V outlines recommendations based on interpretation of model output and discusses future work. Details relating specifically to the diagnostics for MLR and ZINB models are included in the appendices, along with meta-data and other relevant model details.

THIS PAGE INTENTIONALLY LEFT BLANK

# II. BACKGROUND AND LITERATURE REVIEW

The following chapter provides a lens into the Department of Defense's (DOD) motivation for examining the AC enlisted recruiting market. It examines the qualitative measures taken to understand the recruiting environment and the subsequent steps to quantify these data to help the military compete with corporate entities and universities for the nation's youth. It cites both academic studies and survey methods used by several institutions, thereby creating a timeline of efforts by the DOD to model and predict accessions. The literature review considers two academic studies leveraging socio-economic data and profiling the use of the multiple linear regression (MLR), Poisson generalized linear model (GLM), and zero-inflated Poisson regression (ZIPR) models.

## A. BACKGROUND

The Joint Advertising and Marketing Research Study (JAMRS) office has surveyed the U.S. population since the 1970s in order to gain insight into the attitudes of the nation's youth regarding military service. These surveys, published semi-annually in June and December, and largely unchanged since their inception, provide a qualitative description of youth propensity toward the military based on socio-economic factors like "age, school status, educational prospects, employment, employment prospects, race/ethnicity, and geographic location" (Marsh, 2011, pp. 1–2). "Ongoing information on youth attitudes" alone will not help NRC zero in on a recruit-rich environment (Wilson, 1999, pp. 1–1).

While these surveys did not focus specifically on Navy propensity and Navy enlistees, they provided the catalyst for the collection of socio-economic data to gain a better understanding of the impacts independent socio-economic factors play in recruitment of Navy personnel. These surveys provide decision makers insight into the perception of military service in the target population, but limited insight into the socio-economic factors that affect recruit production in an area. With the proliferation of governmentally collected socio-economic data, there is potential to leverage statistical modeling and analysis to provide decision makers greater insights into an area's

recruiting potential. The Navy recruits in all U.S. states and territories. Navy Recruiting Command (NRC) consisted of 952 (as of May 2015) Recruiting Station IDs (RSIDs), which make up the 26 Navy Recruiting Districts (NRD). These NRDs are divided into two regions, east and west, and those two regions together cover NRCs area of responsibility. Figure 1 shows a map highlighting NRCs geographic hierarchy.

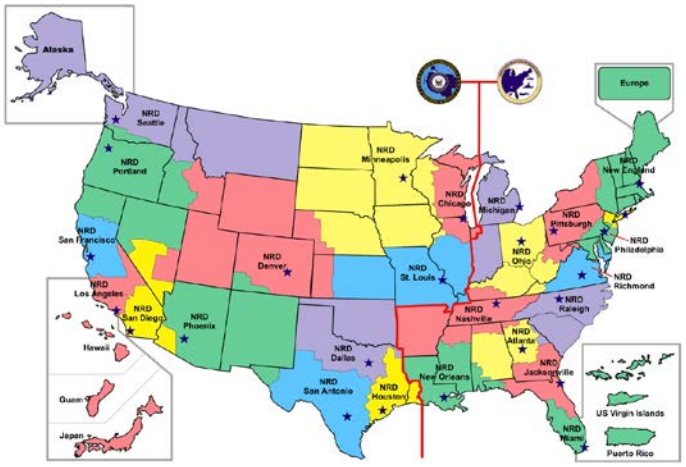


Figure 1. Map of NRC districts and regions (from Commander Navy Recruiting Command [CNRC], 2011).

Between 2011 and 2013, the total number of recruiters decreased from 3,254 to 2,862 (R. Powell, personal communication, May 2015). During this same period, the number of accessions increased from 32,668 to 38,574 (R. Powell, personal communication, May 2015). Since each recruiter is expected to produce a greater number of recruits each year, the identification of those areas with rich recruiting markets is more critical than ever to ensuring that the Navy meets its recruiting mission. Developing statistical models at the station level based on ZIP code-level socio-economic factors can provide the precision needed for stations to better position their recruiters and find the pool of candidates qualified for service in the Navy.

## B. CURRENT PRACTICES

This section describes the qualitative approach to recruiting as well as an example of the quantitative approach (Jackson, 2015). These two methods complement each other.

Surveys sanctioned by the DOD following the termination of the draft eventually became the foundation for other qualitative studies and later quantitative studies to gain a better understanding of why people enlist and the socio-economic factors that should be used in making such predictions.

Following the termination of the draft, many within DOD wondered: "Will the armed forces be able to recruit the necessary number and quality personnel; will this force be relatively representative of the larger society" (Janowitz, 1973, p. 87). In the months leading up to the end of the draft, Janowitz asserted, "the social demography of the armed forces is not predetermined, but it will play an important role in the internal viability of the armed forces and in civil-military relations in the post-Vietnam period" (Janowitz, 1973, p. 87). In an effort to understand the perception of the military by the military-age population of the U.S., survey efforts were undertaken.

In the wake of Janowitz's journal article, DOD conducted Youth Attitude Tracking Studies (YATS) from 1975 until 1999 to gain insight into youth attitudes toward the military and determine the demography of these potential recruits (Wilson, 1999). In the *1999 Propensity and Advertising Report*, a survey, consisting of a 30-minute interview with a nationally representative sample of youth ages 16–24, was conducted, identifying several general demographic trends of youth interested in the military (Wilson, 1999). In general, "propensity: declines with age; declines with increasing educational attainment; was higher for unemployed youth than employed youth; and varies by region" (Wilson, 1999, p. iii). Not only do these surveys show trends with respect to demographics, but also they show general trends with respect to time and current events like war and economic recessions. Following the termination of YATS in 1999, JAMRS initiated similar survey efforts to continue to provide the DOD recruiting enterprise insight into youth attitudes toward military service. JAMRS expanded these survey efforts to provide a clearer understanding of a geographic area's potential to produce recruits by increased sampling and expanded the portfolio of survey instruments to include Media Surveys, Futures Surveys, and Medical Student Surveys. These general tools, which tap into a nationally representative sample of the nation's youth, provide a foundation to guide recruiters toward youth with a high propensity for military service,

but provide little insight into expected accessions from an area and the socio-economic factors that affect recruiting.

### 1. Statistical Analysis

Previous researchers used statistical analyses to:

1. Develop models with predictive power to approximate ZIP code and station-level accessions.
2. Identify the most revealing independent variables in the data set.
3. Develop region-specific models that provide accession goals for NRDs and NRSs.

NRC incorporated a series of analytical tools to aid in assigning resources and assist in identifying areas likely to produce recruits. Analytical tools currently in place within NRC include Standardized Territory Evaluation and Analysis for Marketing (STEAM), Station Market Analysis and Review Technique (SMART), and the Noble Index.

NRC uses the STEAM database as its market research tool for determining a geographic region's recruit potential. Inputs to the STEAM database include "demographic, Navy, All Service Accession Data (ASAD), LEADS, and ASVAB test taker data" (CNRC Publications, 2011, Ch, 3, p. 1). Race, ethnicity, and Test Score Category (TSC) then segment these data and this provides the input for the Navy Recruiting Regions' (NRR) and districts' algorithms to develop a goal matrix (CNRC Publications, 2011). This analytical process was an improvement on the processes of surveys and questionnaires to determine propensity. Analysis of the STEAM database enables NRC to place recruiters in the most recruit-rich environments and set accession goals at the district and station level. WebSTEAM is the current version of STEAM used at the district level. Its database is updated daily, monthly, quarterly, and annually and includes:

1. All Service Contracts (Gross)(ASCD) and All Service Accession Data obtained from the Defense Manpower Data Center (DMDC).
2. Navy New Contract Data (NET) obtained from Personalized Recruiting for Immediate and Delayed Enlistment (PRIDE).

3. ZIP Code Demographic Data purchased commercially.

4. Recruiting Facility Data from Recruiting Facilities Management Information System (RFMIS).

5. Recruiting Personnel Data from the NAVCRUITCOM Personnel Status Report (PSR).

6. High school and College/University data obtained both commercially and from United States Military Entrance Processing Command (USMEPCOM).

7. ASVAB Student Testing Program (STD) data obtained from USMEPCOM.

8. Advertising LEADS data for local and national advertising obtained via National Advertising Leads Tracking System (NALTS) (CNRC Publications, 2011, Ch. 3, p. 2).

With all data input into the model, STEAM reports provide for the user the following production output:

1. Goal Matrix—goals and sub goals by station.

2. Leads ZIP Code Report—ZIP code alignment of NRS given to NALTS to verify current ZIP codes of NRS.

3. Market Share—ZIP code-level demographic and production data for NRS, zone or the entire NRD.

4. District Summary—summary of NRD demographic and production totals or NRS, zone or demographic and production totals for entire NRD.

5. Station/Zone Summary—summary of NRS and zone demographic and production totals (CNRC Publications, 2011,Ch. 3, p. 2).

Number 3, market share, is a proportion of a selected data element indicating the fraction that each station encompasses within a district market. These values are weighted and current practice is “50% weight on the total 17–21 male markets” (CNRC Publications, 2011, Ch. 3, p. 5). This is the foundation for calculating the Recruiter Assignment Factor (RAF) (CNRC, 2011). Figure 2 provides station-manning rules, which vary contingent on outside factors like leadership and market ethnicity.

| STEAM MANNING BUSINESS RULES | |
|---|---|
| **RAF (VALUES)** | **NAVCRUITSTA MANNING** |
| 0-1.8 | 1 - Recruiter<br>*(**Note**: NAVCRUITCOM guidance is that NAVCRUITSTA Territory supporting only 1 Recruiter is a candidate for a Part-time Office (PTO) or possible realignment and consolidation with another NAVCRUITSTA)* |
| 1.81-2.80 | 2 – Recruiters |
| 2.81-3.80 | 3 – Recruiters |
| 3.81-4.80 | 4 – Recruiters |
| > 4.81 | 5 or more Recruiters<br>*(**Note**: NAVCRUITCOM guidance is that NAVCRUITSTA Territory supporting more than four Recruiters is a candidate for realignment and potential new NAVCRUITSTA)* |

Figure 2. U.S. STEAM NRS manning rules (from CNRC, 2011).

These recruiter numbers are then used to assign numbers to the stations; however, once assigned to a station, the recruiters are left to more ad hoc methods and institutionalized knowledge to determine the best ZIP codes from which to recruit. STEAM and webSTEAM provide leaders visibility on data describing their recruiting market, but they do not rely on empirical statistical models to develop their recruiting goals.

## C. LITERATURE REVIEW

The following literature review focuses on previous academic and organizational studies that have identified independent predictor variables and analytical tools used in market research. Due to the breadth of information and goal of this study, our literature review will focus primarily on the quantitative studies conducted in the recent past, specifically, Sandra Jackson's thesis, "Utilizing Socio-Economic Factors to Evaluate Recruiting Potential for a U.S. Army Recruiting Company" and the study by Pinelis et al., *An Analysis of Navy Recruiting Goal Allocation Models.* The goal behind this literature review is to recognize their contributions to the use of socio-economic factors in model development and illustrate how our work can improve on both of the studies listed.

### 1. USAREC Supplied Data and Multiple-Linear Regression

Jackson (2015) sought to use U.S. Army Recruiting Command (USAREC) station-level data to develop monthly multiple linear regression and Poisson Regression models to predict Army accessions. She used data from 250 recruiting companies across a span of four years for a total of 10,323 observations. Jackson's goal was to improve on the tactical segments the Army currently uses. Tactical segments are how USAREC identifies a specific geographic area based on certain demographic characteristics. Jackson used the following socio-economic data in her modeling approach: recruiters, unemployment rate, QMA population, metro (the number of ZIP codes with populations over 50,000), micro (the number of ZIP codes with populations between 10,000 and 50,000), and other (the number of ZIP codes with populations under 10,000). Rather than using variable selection to find the best subset of variables, Jackson tested 16 different permutations of the original variables.

Jackson used four years of data to train her models and conducted out-of-sample 10-fold cross validation using 1,323 randomly selected observations from the original data set. In order to compare both of the model types, Jackson used the Negative Log Likelihood (NLL) statistic to compare MLR and Poisson. Comparing the MLR models, she used the traditional $R^2$ statistic. For both in-sample data and out-of-sample performance tests, her MLR models out-performed her Poisson models with one exception. Conclusions from Jackson's study of Army recruiting are as follows:

1. MLR models out-perform Poisson Regression models when applied to annual company-level recruit production.
2. QMA was the principal driver in predicting the number of Army accessions.
3. If the MLR model was attempting to model the number of Army accessions, then the number of recruiters should be one of the predictor variables.

Based on Jackson's work, we developed a MLR model of annual Navy accessions at the station level. Jackson's data, supplied by USAREC, consisted of only five independent variables commonly collected by the Army recruiting enterprise. Here, we start with 71 potential independent variables gathered from open sources. One of the

goals of this research is to understand the socio-economic factors that influence recruiting. Jackson's analytical tools modeled accessions and recruits per recruiter at the monthly level. We provide national and Navy region (east and west) models of annual accessions using higher fidelity ZIP code-level data, based on Pinelis et al's argument that ZIP code-level data provide a better unit for analysis and are "more responsive to changing demographic needs […] target[ing] specific subpopulations as necessary" (p. 19). The report by Pinelis et al., which makes use of the zero-inflated modeling technique, is highlighted in the following section.

### 2. Center for Naval Analysis's Zero-Inflated Poisson Model

Pinelis et al's study of the Navy's Enlisted Goaling Model (EGM) was the first evaluation of the EGM since the 1990s (Pinelis et al., 2011). The previous EGM was an autoregressive model "designed to determine the supply of eligible recruits and to allocate the recruiting mission to each NRD" (Pinelis et al., 2011, p. 13). The Navy developed the EGM prior to the identification of the need for a diverse force that represents the diversity in the U.S. population and did not account for race, ethnicity, or gender (Pinelis et al., 2011). The Navy's biggest concerns with the use of the EGM were its limited fidelity due to the larger unit of analysis, and the model's lack of support for the Navy's diversity goals (Pinelis et al., 2011). Pinelis et al. attempted to improve on the model's fidelity by developing a ZIP code-level statistical model of Navy accessions. The technique she used was the zero-inflated Poisson regression (ZIPR) model. Chapter IV details the zero-inflated modeling technique in detail.

To address the diversity issue, Pinelis et al. used independent variables describing the demographic characteristics of the population in each ZIP code. The ZIPR models accessions in two steps. First, the model uses a binomial process and a subset of input variables to identify the ZIP codes expected to generate structural zeros (Flynn, 2009, p. 154). Secondly, the model uses another set of input variables to predict the number of accessions from the non-structural zero ZIP codes. The subset of input variables used in the binomial process of the ZIPR model included:

1. Distance to the nearest college or university

2. Size of the college or university
3. Interaction of size and distance
4. Multi-school flag
5. Historically Black College or University (HBCU)
6. Once the ZIPR model identified the ZIP codes with structural zeros, then the Poisson component of the ZIPR modeled the expected accessions from those ZIP codes that did not predict structural zeros. Pinelis et al.'s predictor variables included:
7. Distance to the responsible NRS
8. Demographic data
9. Navy Awareness Index
10. Recruiters
11. Crime Data
12. Veteran Population

The model was developed using five years of data to predict the subsequent year's accessions (e.g., 2006 data to fit a model that predicts 2007 accessions, 2009 data to predict 2010 accessions, etc.). General conclusions regarding the input variables were that veterans ages 17–44 and 85-plus had a positive association with the number accessions. Veterans, ages 45–84, had a negative association with accessions. Recruiters, Navy awareness and crime data all had positive association with accessions, but distance to the responsible NRS had a negative association (Pinelis et al., 2011).

Pinelis et al. used mean absolute deviation (MAD) and the rate of false positives as metrics to evaluate her models' performance. MAD is the mean of the absolute values of the residuals, across all ZIP codes. For their 2010 model, this was .943, meaning that her prediction was off, on average, by one person per ZIP code (Pinelis et al., 2011). Their second performance metric, false positives, was the number of times a model predicted non-zero accessions for a ZIP code that actually had zero accessions. The model correctly identified just over 55% of the zeros. Because of this false positive rate, Pinelis et al. (2011) checked the accuracy of the MAD from only those ZIP codes that yielded recruits. Their updated MAD was .533, significantly better than previously

calculated with the binomial check portion of the ZIP model. Pinelis et al.'s Poisson regression model performed well, but the full model's performance could be improved by lowering the false positive rate when identifying ZIP codes that are actual zeros.

This research expands on Pinelis et al.'s work by incorporating IRS ZIP code-level data to determine if there exists a relationship between financial profiles and accessions. Furthermore, efforts were made to improve the capability and accuracy of predicting structural zeros, which could have far-reaching impacts for NRC because accurate predictions of ZIP codes with structural zeros could result in cost-savings measures by the Navy (e.g., closing stations, realigning stations, etc.) (CNRC, 2011).

# III. DATA COLLECTION AND CLEANING AND CONSTRAINTS, LIMITATIONS, AND ASSUMPTIONS

## A. DATA SETS

All original data sets were collected from open source websites, including IPEDS, IRS, FBI, and the U.S. Census Bureau. This differs from NRC, which pays for commercially available data sources (CNRC, 2011). Data were selected based on availability, accessibility, quality, and the previous research in this area. Data were collected for the years 2011 through 2013. A three-year period was chosen because the average tour length of a Navy recruiter is approximately three years.

Data were collected at the ZIP code-level in order to aggregate up to the Navy Recruiting Station (NRS) level, since stations are collections of ZIP codes. A master ZIP code file was built containing 32,980 ZIP codes (United States Census, 2010). The population data used throughout the data set was from 2010 ZIP code census data, since minor adjustments to ZIP code boundaries are made annually (ZIP Boundary, 2014). Only 32,980 ZIP codes inside the 50 U.S. states and DC were considered.

Since some data sets were only available at the county level, data preparation required a mapping from county to ZIP code-levels. For example, crime data from the FBI was only available at the county level. A ZIP to FIPS (Federal Information Processing Standard) county code crosswalk used to match the ZIP codes to the appropriate FIPS. The cross-reference file for 2010 ZIP codes was available at the U.S. Census Bureau site. Building the proportional crosswalk key for each ZIP code involved using the 2010 census ZIP code populations for all 32,980 ZIP codes along with their respective county and state. For the purposes of this description, Federal Information Processing Standards (FIPS) code and county are used interchangeably. ZIP codes were grouped by the convention, “state.county” key, and the sum of all ZIP codes within that county were summed and used as the denominator for the population of each ZIP code to determine the proportion it made up of that county. A data frame with all ZIP codes, their respective “state.county” key, and their proportions was the resulting output.

Once we made the correct cross-reference and determined the proportion each ZIP code made up of its county, we multiplied that proportion by the observation for that category in that county. If the rate for that county was available via open source (e.g., unemployment rate), then the rate was applied to all the observations in that county. R code scripts used for this conversion are in Appendix E.

Since the lowest unit of analysis was the ZIP code, the majority of the data sets were available at the ZIP code level, and the remainder was easily mapped to the ZIP code level, the full data set was first prepared at the ZIP code level. The next level of analysis desired was at the station level. Since stations consist of a collection of ZIP codes, with each ZIP code being assigned to only one station, a methodology was required to aggregate the data to the station level. Depending on the data, the mean, median, maximum, or sum across all ZIP codes assigned to the station was used at the station level. The method chosen for each independent variable will be discussed later with each category of data.

Below is a description of the data used and how each independent variable was cleaned and aggregated. Data were partitioned into six categories: military influence and recruiter workload, crime, population characteristics, economic stability, education opportunities, and veteran population. Across these six categories a total of 71 independent variables were prepared to support the analysis. The following sections explain the rationale for each category, the predictor variables selected, and the methodology behind manipulating the original data sets to obtain the data used in the analyses.

### 1. Military Influence and Recruiter Workload

Military influence on youth comes from multiple sources including friends, parental influence, advertising, a sense of civic duty, and proximity to military recruiting stations and bases (Wilson, 1999). Recruiter numbers provide one way to quantify influence; according to the Naval Postgraduate School (NPS) thesis by Captain Taylor Williams, he determined that the "average number of recruiters over the last 12 months at each NRS […] is the most important variable in all models" (Gibson, 2009; Williams,

2014, p. xviii). Recruiter workload, interpreted by distance from NRS to Military Entrance Processing Station (MEPS), was also identified in Williams (2014) as an important factor in the number of accessions.

By identifying the effect of factors that NRC can control directly like station alignment, number of recruiters, recruiter to QMA ratio, and distance to NRS, NRC has the ability to change how it utilizes these resources and increase efficiency. Below, we describe each variable within the military influence and recruiter workload category, and the method for arriving at the final input variable.

### *a. Recruiters*

NRC provided recruiter data for the years 2011–2013. Data refer to average number of recruiters assigned to the station during each year. The original data provided by NRC were the average number of recruiters per station for each year. To create a station level model , using a function in R, we summed up all the averages for each ZIP code by station. We determined the average number of recruiters at each station across the entire year. Input data to the model at the ZIP code-level was so small compared to the other variables in the model that scaling by a factor of 100 was required (S. Buttrey, personal communication, September 2015).

### *b. Recruiter to QMA*

The ratio of recruiters to QMA determined a recruiter's workload. Woods and Poole (W&P) QMA data, alone, were a predictor variable and will be explained in greater depth in the following chapters (Woods & Poole, 2015). In order to get this ratio, we divided the total number of recruiters for each year, 2011–2013, at each station by the number of QMA within that station's area of responsibility. This ratio method was applied at the ZIP code-level. Of the 98,940 observations (32,980 x 3 years of data), approximately 11,000 had zero QMA. This suggests that approximately 3,700 ZIP codes, or 11%, have zero QMA. To avoid errors in R, those ZIP codes were assigned a ratio of zero (R Core Team, 2013).

### *c. Distance to NRS and Number of Stations Within 50 Miles*

These two predictor variables go together because they came from the same source data and PYTHON code (Rossum, 1995). A distance calculator we coded in PYTHON, coupled with source code written by U.S. Army Captain Sandra Jackson of the University of Texas, determined the distance between each ZIP code centroid and its nearest NRS (Jackson, 2015). The number of NRSs within 50 miles of each ZIP code centroid was another measure of convenience, captured with the same PYTHON code. Driving long distances for recruiters and leads is inconvenient. Recruiters spend an inordinate amount of time doing this. Once the distances were determined for each ZIP code, the median distance was used at the station level. Derivation of the station level variable, "minimum distance to a recruiting station from the ZIP code centroid" is as follows (R code and PYTHON scripts used for this conversion are in Appendix E):

1. Sort ZIP code distances by station ID
2. Calculate the median distance of all ZIP codes for each station ID

When deriving the "number of stations within 50 miles" at the station level, we used the mean of all ZIP codes within that station. Derivation of the station level variable, "number of stations within 50 miles" is as follows (R code and PYTHON scripts used for this conversion are in Appendix E):

1. Sort ZIP code distances by station ID
2. Calculate the mean station count for all ZIP codes in each station ID boundary

### *d. Station Area*

We received station area from the 1October2014 NRC station area data. Using the 2014 station boundary area, we are assuming the station area boundaries for 2011–2013 are all identical to 2014 regardless of change in station ZIP code alignment. Since area data were not available at the ZIP code-level, they were only used in the MLR models. This provides a measure of the workload of the recruiter.

### 2. Crime

Crime, as determined by Malone (2009) and cited by Pinelis et al., "has shown to be an important variable for predicting enlistments" (p. 23). Furthermore, a related Army study concluded, "annual property crimes per capita were positively related to accessions such that ZIP codes with more property crimes tended to yield more accessions" (Gibson, 2009, p. 25). This same study also noted that violent crime did not affect accessions. In light of this study, we acquired the FBI's Uniform Crime Reports (UCR) to incorporate into our research (Federal Bureau of Investigation, 2015). As stated in the assumptions paragraph, data were not available at the ZIP code-level, so a crosswalk data set was required to weight the crimes by population within each ZIP code. This method differs from the method Pinelis et al. used. They applied rates from the U.S. Census Bureau state level crime rates to each ZIP code in the state. Arguably, the weighted approach provides more fidelity to each ZIP code with respect to crime.

Requirements for reporting crime to the UCR were strict and data that were not provided in accordance with those standards were not used in the UCR. A list of footnotes for each year is provided from the UCR and can be viewed in Appendix B. For this reason, the following states did not report any data for either one or all three years used in the study: (Alaska, Connecticut, District of Columbia, Hawaii, Massachusetts, and Rhode Island). Table 4 provides the fractions of missing data for all ZIP codes and all years combined.

In the states that did not report, recall Figure 1 and note that none of the states listed is partitioned by a NRD and that each NRD is the aggregate of its NRSs and those stations are the aggregate of ZIP codes. Crimes in these states are estimated by:

1. Calculating the median of all reported violent crimes and nonviolent crimes for every ZIP code within each NRS boundary and applying the estimates to the unreported ZIP codes within the station boundaries for each respective year.

2. Calculating the median of all reported and estimated violent crimes and nonviolent crimes for every ZIP code within each NRD boundary and applying the estimates to the unreported ZIP codes within the district boundaries for each respective year.

Following the second iteration of estimates for ZIP codes at the district level, all 32,980 ZIP codes contained an estimated or reported value. Noting Table 1, the median value was used because of the large variance when compared to the mean, for both violent and nonviolent crime.

Table 1. Non-zero crime statistics for reported crimes from the uniform crime reports, all ZIP codes.

| Statistic | Violent Crime | Nonviolent Crime |
|---|---|---|
| Mean | 8.67 | 14.61 |
| Median | 1.43 | 70.13 |
| Variance | 508.57 | 33572.16 |

### *a. Violent and Nonviolent Crime*

Since the original crime data were provided by the FBI at the county level, divided into metropolitan and non-metropolitan counties, we developed keys to identify each unique entry, under the assumption that no one state had two counties of the same name. Some of the counties listed the reporting authority (e.g., Fairfax County Police Department) so the "County Police Department" was stripped from the data. The data sets also provided a breakdown of every type of crime, which fell into the main two categories of violent crime and nonviolent crime. These other crimes were not used, only the totals. With "state.county" keys complete, next we stripped the trailing and leading spaces, so that the keys matched the FIPS code to ZIP code proportion crosswalk keys. Due to the footnotes, we removed several superscripts from the county names, as well. At this point, we had three columns to work with in the crime data set, keys, violent crime, and nonviolent crime.

The majority of the changes made applied to all three years; however, there were some unique changes that applied only to specific years. For these cases, a "changers" file was created and run in R to make the changes throughout the entire crime data frame to ensure all keys in crime matched all keys in the proportion data frame (R Core Team, 2013).

Table 2. Portion of the output file with the proportions each ZIP code makes up of its respective county. For example, ZIP code 36003 makes up 2.3% of the total population of FIPS Code 1001 (Autauga County); therefore, we make the assumption that 2.3% of the crime that occurs in Autauga County. If there are 100 crimes reported in Autauga County, 2.3 of them are in ZIP Code 36003.

| FIPS Code | ZIP Code | Weighted Proportion | County | State 1 | State 2 |
|---|---|---|---|---|---|
| 1001 | 36003 | 0.023 | Autauga | AL | ALABAMA |
| 1001 | 36006 | 0.016 | Autauga | AL | ALABAMA |
| 1001 | 36022 | 0.155 | Autauga | AL | ALABAMA |
| 1001 | 36051 | 0.027 | Autauga | AL | ALABAMA |
| 1001 | 36066 | 0.226 | Autauga | AL | ALABAMA |
| 1001 | 36067 | 0.319 | Autauga | AL | ALABAMA |
| 1001 | 36091 | 0.043 | Autauga | AL | ALABAMA |
| 1001 | 36703 | 0.161 | Autauga | AL | ALABAMA |
| 1001 | 36749 | 0.013 | Autauga | AL | ALABAMA |
| 1001 | 36758 | 0.017 | Autauga | AL | ALABAMA |

Once all the keys matched between the FIPS crime data frame and the ZIP code proportion data frame, we ran a loop in R that matched the key from the crime data to the key in the proportion data and multiplied the crime by the ZIP codes' proportion of its respective county (R Core Team, 2013). The final output was the weighted proportions of violent crime and nonviolent crime at the ZIP code-level. R code scripts used for this conversion are in Appendix E.

### 3. Population Characteristics

General population characteristics helped to achieve a clearer understanding of a demographic. We considered only three variables in this category: QMA, 2010 census population data, and population density (Woods & Poole, 2015). QMA, supplied by Woods and Poole Economics, tracked the qualified pool of the population available for military service. Specifically, it was "the total 17–24 year old population, excluding institutionalized and those in military service, unauthorized immigrants, and non-high school diploma graduate (HSDG) not enrolled in high school or an equivalency program" (Pinelis, 2011, p. 17). The population data was a function of the size of the pool from

which recruiters were seeking those QMA. All three variables could be combined into recruiter workload as they were functions of recruiter effort, but because the data were not specific to NRC, we combined them into general population characteristics.

### 4. Economic Stability

Historically, wealth has been a generally regarded hallmark that distinguished those who go to college from those who enlist (Kleykamp, 2006). It was an accepted norm that high school graduates who came from affluent families had more opportunities than those who came from more modest backgrounds (Kleykamp, 2006). Economic data were pulled from the IRS SOI website. Because recent studies, Jackson (2015) and Pinelis (2011), did not exhaust or even incorporate economic variables into their models, we saw an opportunity to test our hypotheses and make full use of the open source economic data available. We used the Internal Revenue Service (IRS) individual statistics of income (SOI) data set, available at the ZIP code-level, to examine the relationship between economics and Navy recruit production (Internal Revenue Service, 2015). This data set included 77 financial variables that provide information from every household that filed a tax return. Predictor variables were derived from six of the 77 financial variables. R code scripts used for this conversion are in Appendix E.

#### *a. Mean Unemployment Rate*

Unemployment rate is the ratio of the number of those who are unemployed to the size of the entire workforce. Data were obtained at the county level for each month over the three-year span our research covered. The rate for each county was applied to each of the respective ZIP codes in its county. Because this study was designed to develop annual models, the mean from each year was used.

#### *b. Standard Deviation of Unemployment Rate*

In an attempt to identify the stability of the job turnover rate and determine whether it had an effect on enlistments, we considered the standard deviation of the monthly unemployment rates. Using the original monthly data on each ZIP code, we determined the standard deviation for each year, postulating that a high variance would

indicate some instability with the market and result in more accessions for those youth seeking more stability in the workforce.

### *c. Average Adjusted Gross Income (AGI)*

AGI was a metric available from the IRS's individual SOI data. To develop this metric, we divided the sum of all AGIs in the ZIP code by the total number of returns and multiplied the quotient by 1,000 to obtain the average AGI in thousands of dollars.

### *d. Unemployment Compensation*

Previous studies, including those of Pinelis and Jackson, have included the metric of mean unemployment rate as a way to identify a market with few job opportunities. Because of the inherent phenomena hidden in unemployment rate (i.e., persons who have not sought employment in the previous four weeks or those who were not in the labor force), we required another variable to identify the impact of joblessness (United States Department of Labor, 2014). Not working was not enough to be considered unemployed, so another metric to determine the rate of joblessness was required for our analysis. Unemployment compensation was the other metric to measure the jobless rate of a specific area and an incentive one would have to not enlist. The metric was derived by taking the ratio of the number of returns with unemployment compensation, regardless of amount, to the total number of tax returns for that ZIP code.

$$compensation = \frac{n_{compensation}}{r_{returns}} * 100 \tag{3.1}$$

The targeted metric was the number of people receiving compensation for being unemployed, not the amount of dispensation paid. Input data to the model at the ZIP code-level was so small compared to the other variables in the model that scaling by factor of 100 was required (S. Buttrey, personal communication, September 2015).

### *e. Pensions and Annuities as a Fraction of AGI*

In an attempt to model areas with a high number of retirees and / or wealth, we wanted to consider an economic metric that was unique to those two categories. Many retired and wealthy households receive a large portion of their AGI from pensions and

annuities. By eliminating areas with a high concentration of retirees and wealthy households, recruiters can focus their efforts on other areas. One particular area of Florida, known as Sun City Center, ZIP code 33573, is a well-known retirement community. The fraction of the incomes in that ZIP code that come from pensions and annuities was .48. The spread nationally across all ZIP codes ranges from 0 to .77 with a median of .13. This metric was derived by taking the total dollar amount of returns with taxable pension and annuity payouts and then dividing it by the total dollar amount of all returns in the ZIP code.

### *f. AGI Categories and Six Figure Incomes*

Believing a relationship exists between income and accessions, we partitioned IRS income data into 7 categories to examine the affects. Few studies that we examined in our literature review showed a quantitative relationship between income and recruitment potential; yet, a large number of qualitative studies have identified a positive relationship between membership in the lower income brackets and accessions. Partitioning the population into its respective income brackets was a way to identify and target a particular market, as there were several unique demographic characteristics that follow each bracket. AGI, by ZIP code, was divided into six categories as seen in Table 3. Based on the theory that enlisted accessions generally come from the lower income brackets, we would expect to observe a negative relationship between category number and accessions.

Table 3. Categories and ranges of Adjusted Gross Income

| Category | AGI Range ($) |
|---|---|
| 1 | 1–25,000 |
| 2 | 25,001–50,000 |
| 3 | 50,001–75,000 |
| 4 | 75,001–100,000 |
| 5 | 100,001–200,000 |
| 6 | > 200,000 |

AGI Categories 5 and 6 were combined to form one category of six-figure incomes. We postulated that earners in the top 25% ($89,125 and higher) would value the worth of a college education for their children and have the resources to send them to a university *(New York Times*, 2012).

### 5. Education Opportunities

Education data were taken from the Integrated Postsecondary Education Data System (IPEDS) website. The criteria used to extract university data from IPEDS were four-year degree-granting, public and private, not-for-profit universities. The list provided 2,339 schools that met the aforementioned criteria. We wanted to keep the higher education factor as traditional as possible excluding for-profit and vocational schools, which would increase the number of schools by more than half and possibly skew our results. We wanted to focus on the traditional 17–24 year old student, which is where the Navy focuses its recruiting efforts (CNRC Publications, 2011).

#### *a. Division One Universities*

Pinelis included similar data in her report. Our data differs in that we singled out Division I universities and dedicated two variables strictly to those 351 schools (Athnet, 2015). Division I universities are those schools that "generally have the biggest student bodies, manage the largest athletic budgets and offer the most generous number of scholarships" (NCAA, 2015). With that data, using PYTHON and the distance formula in Appendix E, we determined the distance to the nearest Division I university from every ZIP code centroid and counted the number of Division I universities within 50 miles of every ZIP code centroid. It was a method identical to determining nearest NRS and number of NRSs within 50 miles. As the distance between Division I schools and a particular ZIP code increases, the influence of academia should decline, thereby increasing the number of accessions. Also, in areas with a high density of schools, for example, the northeast, where some ZIP codes had 10+ Division I universities within 50 miles, there could be more propensity toward a university education.

### *b. Total Schools and High Enrollment*

Division I universities are a large attraction and driving force with which the Navy must compete for youth, but they only make up roughly 15% of not-for-profit public and private four-year universities in the country. Curious to see how the ZIP code and station density of *all* schools affected accessions, we computed the total number of universities in each ZIP code. To understand the effect of school size, we captured high enrollment as a variable, recording the enrollment of the most populated school in that ZIP code. High, in this case, means the largest student body. Of note, four-year universities are not established overnight, so for modeling purposes, if significant, this was a variable that could be forecasted a few years out for goaling purposes.

### *c. University Population Categories One Though Five*

Schools were divided into five population categories, under 1,000, 1,000–4,999, 5,000–9,999, 10,000–19,999, and greater than 20,000. This portion of the data set was identical to that of Pinelis and warranted another look to determine if there was a relationship between accessions and the number and size of the schools in a ZIP code and station.

### *d. University Tuition*

Tuition is perhaps the greatest barrier to achieving a college education. Citing a June 2000 study by Susan Dynarski, who concluded that for each $1,000 in subsidies, the college attendance rate of middle and upper-income youth rose by four to six percentage points, we were motivated to explore the relationship between tuition, student aid, and accessions. If Dynarski's results were accurate, then one would expect to see a positive relationship between increased tuition and accessions and a negative relationship between increased student aid and accessions. Percentage of student aid was also considered; however, data were sparse, so the variable was dropped altogether.

Because of the high tuition cost of college, the Navy has programs to help students pay for college, like the post 9-11 Government Issue (GI) Bill. With rising college tuition costs and an overwhelming incentive from the Navy, we postulate that in areas with higher tuition, high school graduates would be more inclined to enlist. In-state

and out-of-state tuition was considered. These data made the actual analysis difficult as schools with zero for tuition indicated that there was no tuition, but in reality, it meant there was simply no school in that ZIP code. Concerned that this would skew our analysis, the variable was disregarded. Future work might include a flag that would indicate to ignore the variable if the total number of schools was less than one.

### 6. Veteran Population

The general consensus from previous studies on the topic was that veteran influence was positive in affecting accessions. Data on veteran status were compiled from the 2011–2013 three-year American Community Survey (ACS) county level data (United States Census Bureau, 2015). Since these data were only available at the county level, we calculated a weighted average based on population in the ZIP codes that made up each county. The method used for the crime data was also used for the veteran data. Veteran data were divided into sex and age category. The five age categories were 18–34, 35–54, 55–64, 65–75, and older than 75. Total veteran population and total male and female population were also used. One explanation, which was supported by multiple survey data and studies, was that veterans of "popular wars" would have a positive influence on accessions and those veterans of generally "unpopular wars" would have a generally negative impact on accessions (Pinelis et al., 2011).

In the zero-inflated model development phase of the research (see Chapter IV), we determined that the degree of multi-collinearity among partitioned veteran data was too high, so the categorized data were removed. The only veteran variable used in the zero-inflated models was total veterans and male veterans ages 18–34. Male veterans ages 18–34 identify the impact of peers who have fought in the current Middle East conflict and whether or they have a positive impact on recruiting.

## B. CONSTRAINTS, LIMITATIONS, AND ASSUMPTIONS

### 1. Constraints

Due to the non-availability of monthly data, our models were constrained to predicting annual accessions only. Below is a summary of the fitted models:

1. Annual regional– east, west, and national-level models using ZIP code-level data aggregated to the station-level with accessions as integer variables

2. Annual regional– east, west, and national-level models using ZIP code-level data with accessions as integer variables

The second constraint was the modeling of Active Component (AC) enlisted accessions only. Officer and reserve accessions inherently have their own set of unique modeling characteristics and represent only a small fraction of NRCs recruiting mission. Enlisted accessions comprise the Navy's largest market and fraction of annual accessions (Pinelis, 2011).

### 2. Limitations

Our model predictions are based on input variables (e.g., recruiters, AGI, total veterans, etc.). In order to predict future accessions for year 201X, the modeler must know the number of recruiters, AGI, total veterans, etc. for year 201X. In this sense, the model acts as a "fantasy draft" tool where modelers can use projected or known input data to project an outcome of accessions (Pinelis et al., 2011, p. 66). The modeler can add or subtract recruiters, change station areas, include more or fewer veterans, etc. and then make real-life policy decision based on output. This provides modelers the ability to see how changes in the market affect their accessions.

### 3. Assumptions

Throughout the course of research, the study determined that certain assumptions would have to be made regarding ZIP code alignment because of the changes occurring with station boundaries and populations. For calculation purposes, certain annually changing variables within the study must remain constant. For example, in 2011 a station contained a fixed number of ZIP codes. However, the following year's data showed the same station adding or subtracting ZIP codes from its AOR, contingent upon the needs of the Navy. These boundaries change only slightly each year; a small variation, < 1% of the ZIP codes, makes such an insignificant change that updating boundaries each year was not warranted (CNRC Publications, 2011).

NRC does not recruit from ZIP codes not containing a population. Data collected from a ZIP code having zero population was useless and a waste of resources. A master list consisting of 32,980 ZIP codes was created and served as the basis for all three years of data gathered, regardless of a change in alignment.

Data were not always available at the ZIP code-level, so assumptions regarding county level data had to be made. Assuming that the independent variable collected was a function of population, a weighted proportion based on ZIP code population was used to determine how to distribute the data across ZIP codes.

Inherent in most data analysis is the problem of missing data. Assumptions had to be made to estimate these missing observations. If data were not available for a certain ZIP code, then the distribution of the available data was determined through analysis of available data. Once the distribution was determined, an estimate (e.g., mean, median, maximum, etc.) could be used for the missing values, denoted as NAs. The estimate was derived from available data at the next higher unit in the NRC hierarchy (i.e., ZIP code, station, District, Region, NRC). For example, if three ZIP code-level observations were missing data, then the estimates for those NAs would be some estimate of the observations in that station. At the ZIP code-level for all three years combined, the following variables contained NAs as referenced in Tables 4 and 5.

Table 4. Missing data (NA) from the original data sets for each observation (98,940). Considering every ZIP code across all three years of observed data, this table shows the number of observations missing and the fraction of the entire data set.

| Variable | NAs/98,940 | % Missing Data |
|---|---|---|
| average number of recruiters per year | 6 | <1 % |
| QMA | 9 | <1 % |
| population density | 9,429 | 9.5 % |
| violent crime | 27,088 | 27.4 % |
| nonviolent crime | 23,813 | 24.1 % |
| mean unemployment rate | 3 | <1 % |
| standard deviation of unemployment rate | 3 | <1 % |
| average AGI | 10,460 | 10.6 % |
| unemployment compensation | 10,460 | 10.6 % |
| percent of returns which include taxable income | 10,460 | 10.6 % |
| pensions and annuities as a fraction of AGI | 10,460 | 10.6 % |
| mean of all AGI categories, 1–6 | 13,842 | 13.9 % |

Table 5. Missing data from the original data sets. Tuition data were unique to the other variables because not every ZIP code included a university; so to say that an observation was missing a data point when a school was not present would be a misnomer. A datum was NA only if the value was missing and a school was present in that ZIP code.

| Variable | NAs/5,262 | % Missing Data |
|---|---|---|
| in-state tuition | 1,264 | 24.0 % |
| out-of-state tuition | 1,264 | 24.0 % |

Because this technique was used consistently throughout the entire data set, it was determined to be the best way to handle NAs when compiling the data to the station level. Also, because of the nature of certain data, it did not make sense to use one estimate operation to compile all data. So for example, for the missing crime data, the median was used when replacing the missing values. However, for the aggregation to the station level, the data were summed to show a count of crimes within a specific station, rather than the median.

## C. TRAINING AND TEST DATA SETS

Following compilation of all data, two subsets were created. An in-sample data subset (2011) was created from which to fit the models. Another out-of-sample data subset (2012) was created as the test set and used in the performance test phase.

# IV. ANALYSIS

This section provides an overview of the statistical methods used in this research and documents the resulting models.

## A. MODELING ANNUAL STATION LEVEL ACCESSIONS

### 1. Multiple-Linear Regression Model Selection

For data aggregated to the station level, the response was the total number of annual accessions for a respective station. Based on previous research and an initial analysis of the predictor variable trends, we decided to use the MLR to train a model to predict accessions and determine the most statistically significant independent variables (Jackson, 2015).

### 2. MLR Model Development and Variable Selection

#### *a. Multiple-Linear Regression Model*

Multiple-Linear Regression is a modeling technique used to describe a linear relationship between a dependent response ($y_i$) variable and multiple independent explanatory variables ($x_{in}$) using ordinary least squares (OLS) estimates (Faraway, 2006). The linear model takes on the form :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in} + \varepsilon_i \tag{4.1}$$

Variables are defined as follows:

$y_i$: i$^{th}$ observed value of the response

$\beta_n$ : coefficient of the regression variable $x_{in}$

$x_{in}$ : observed value of the regression variable $x_{in}$ corresponding to $y_i$

$\varepsilon_i$: error term, which introduces randomness into the model

Indices are defined as follows:

$i$: number of observation

$x_{in}$ : the value of the independent variable used in the MLR corresponding to $y_i$

The goal in developing these particular models was to understand the relationship between each independent variable and the response and provide a model to estimate annual Navy accessions at the station level. The criteria used in determining goodness-of-fit and comparing like models was the adjusted standard error of the estimate and the extent to which the model appeared to meet the five assumptions of an MLR (Poole, 1970):

1. The error term ($\varepsilon_i$) has a mean equal to zero.
2. The error term ($\varepsilon_i$) has a constant variance ($\sigma^2$).
3. The errors are normally distributed.
4. The errors are uncorrelated.
5. The relationship between the response ($y_i$) and the independent variables is correct.

Assumptions 1–] are assumptions regarding the error terms in the model Assumption 5 is a model assumption. Plots validating each assumption are available in Appendix C. The adjusted $R^2$ statistic was used in lieu of $R^2$ due to the initially large number (71) of independent variables to determine goodness-of-fit. Adjusted $R^2$ provides a measure of the amount of variance in the response accounted for by the model and is adjusted to consider the number of terms in the model (Faraway, 2006; R. Silvestrini, personal communication, 2014). Equation 4.2 provides the definition of adjusted $R^2$ (Faraway, 2006):

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} \tag{4.2}$$

Variables are defined as follows:

$RSS$: residual sum of squares

$TSS$: total sum of squares

$n$: number of observations

$p$: number of independent variables in the model

### *b. Development and Variable Selection*

For model development, we began variable selection by fitting the response variable, accessions, to 71 independent variables. Depending on the model, following validation of the assumptions, the response was either square root (national and east regions) of accessions or cubed root (west region model) of accessions. The model development process consisted of multiple systematic reviews using a series of statistical tests and criteria to select the best-fitting subset of model variables and variable transformations to improve the goodness-of-fit while minimizing multi-collinearity and avoiding overfitting. The dependent variable in our model was number of accessions. Statistically significant variables are those with a p-value < .01, unless otherwise noted.

Overfitting occurs when the complexity of the model is too high (i.e., the number of independent variables used to fit the model) for the size of the training set (i.e., the number of observations) used to fit the model (Faraway, 2006). Multi-collinearity is the case in regression modeling when independent variables are correlated, making "interpretation of regression coefficients more difficult" (Faraway, 2006). Correlation matrices and VIF tables are available in Appendix C.

In order to provide a model with the greatest explanatory power and the lowest number of variables possible, systematic variable selection was used. Using the dropterm() function from the MASS package in R, variables were systematically eliminated from the model (Venables & Ripley, 2002). This approach was used for all MLR models. Once a model with good explanatory power and low standard error was achieved, all linear model assumptions were checked and transformations were made, if required. Once all assumptions were met or checked, then a final review using the dropterm() function was performed (Venables & Ripley, 2002). A summary table of results and adjusted $R^2$ values is provided in Table 8. The following list represents the sequence of steps in the variable selection process:

1. Starting with a lm (linear model) object containing all independent variables, the dropterm() was run, the least helpful variable was determined using BIC as the selection criterion, and then it was removed (Venables & Ripley, 2002). This process was repeated multiple times until the incumbent model was the best possible model with given data.

2. Check for outliers and remove (Faraway, 2006, p. 16).

3. Run dropterm() (Venables & Ripley, 2002).

4. Create correlation matrix using remaining independent variables. Crosscheck with Variance Inflation Factors (VIF). Using VIFs and the correlation matrix, manually remove, one at a time, those with VIFs > 4. (Gibson, 2009).

5. Run the regsubsets() function with the incumbent fitted model to verify with the Mallow's $C_p$ statistic that the incumbent cannot improve with current data (Lumley, 2009). Regsubsets() permutes the model variables and provides summary statistics for each permuted model (Lumley, 2009).

6. Diagnostics of the five assumptions of linear models conducted (Poole, 1970; R. Silvestrini, personal communication, 2014).

7. Incumbent fitted model validated using the predict() function and out-of-sample data (R Core Team, 2013). Summary statistics collected.

The dropterm() function in R allows modelers to exhaust single term deletions from the incumbent model and compare the candidate models to the incumbent (Venables & Ripley, 2002). See Table 6 for an example of the dropterm() function (Venables & Ripley, 2002).

Table 6. An example of the dropterm() output showing the improved BIC by removing the parameter female veterans ages 18–34 variable from the incumbent model, identified by <none>. In this case, the modeler would want to remove the female veterans ages 18–34 variable. Conversely, removing average number of recruiters per year would result in the most severe model degradation.

| Independent Variable | Df | Sum of Sq | RSS | BIC | F Value | Pr(F) |
|---|---|---|---|---|---|---|
| female veterans ages 18–34 | 1 | 0.400 | 96.213 | –2120.554 | 3.931 | 0.048 |
| <none> | NA | NA | 95.812 | –2117.658 | NA | NA |
| mean unemployment rate | 1 | 1.768 | 97.580 | –2107.145 | 17.363 | 0.000 |
| unemployment compensation | 1 | 1.904 | 97.717 | –2105.818 | 18.702 | 0.000 |
| AGI < $25,000 | 1 | 2.413 | 98.225 | –2100.887 | 23.697 | 0.000 |
| AGI > $200,000 | 1 | 3.444 | 99.256 | –2090.966 | 33.824 | 0.000 |
| number of NRSs within 50 miles | 1 | 6.860 | 102.673 | –2058.816 | 67.379 | 0.000 |
| male veterans ages 35–54 | 1 | 13.175 | 108.988 | –2002.115 | 129.396 | 0.000 |
| average number of recruiters per year | 1 | 24.372 | 120.184 | –1909.213 | 239.362 | 0.000 |

The methods of comparison used were the BIC and the F-statistic. BIC imparts a greater penalty for extra observations, which is why we preferred it to Akaike

Information Criterion (AIC). AIC penalizes less with the larger models and allows more variables into the model, potentially leading to an overfit situation. BIC in Table 6 refers to ln($n$), which in the dropterm() function utilizes the BIC for model comparison. From equation 4.3, one can see that as the number of observations $p$ and number of model independent variables $k$ increases, the penalty on the BIC score is greater.

$$\text{BIC} = -2\ln\hat{L} + k\ln(n) \tag{4.3}$$

where $\hat{L}$ is the maximized value of the likelihood function of the candidate regression model.

The F-test identifies whether dropping the variable improves the fit, when compared to the incumbent (Faraway, 2006, p. 14). Using a hypothesis test:

- Null hypothesis: if the independent variable is removed, then it will make no statistically significant improvement in the model fit.
- Alternative hypothesis: removing the variable will improve the fit.

Using a p-value of .01, elimination of insignificant variables was made when comparing the incumbent to the candidate model (Faraway, 2006, p. 14)

Following the first review, a check for outliers was conducted to see if any observations interfered with the variable selection (Faraway, 2006). If outliers were present, they were removed (Faraway, 2006). This process was repeated until the number of independent variables in the model was smaller than 15.

Removal of some variables might not achieve a statistically better fit, so a test should be conducted to determine if the candidate model with the removed variable was a better fit than the incumbent (Faraway, 2006; R. Silvestrini, personal communication, 2014). Next, a review was conducted and a correlation matrix was built from the remaining variables in order to check the level of multi-collinearity present in the model. Variance inflation factors (VIFs), which measure the variance inflation of the regression coefficients, were also checked for values greater than 4 (Gibson, 2009). VIF is interpreted as the increase in the standard error of the coefficient by the square root of the factor, when in the presence of all other variables in the model (Princeton University Library, 2015). Table 7 provides all VIFs for all models and variables.

Table 7. Variance Inflation Factors (VIFs) for all MLRs

| Independent Variable | National | East | West |
|---|---|---|---|
| average number of recruiters per year | 1.32 | 1.46 | 1.31 |
| number of NRSs within 50 miles | 1.30 | 1.62 | NA |
| 2010 population | NA | NA | 2.68 |
| mean unemployment rate | 1.56 | NA | NA |
| unemployment compensation | 1.45 | 1.07 | NA |
| AGI < $25,000 | 2.01 | 2.16 | NA |
| AGI > $200,000 | 1.58 | 1.96 | 1.63 |
| number of universities within 50 miles | NA | NA | 1.23 |
| total veterans | NA | 2.27 | NA |
| male veterans ages 18–34 | NA | NA | 1.72 |
| male veterans ages 35–55 | 1.86 | NA | NA |

Mallow's $C_p$, a summary statistic provided by the regsubsets() function in the leaps package, was another method of determining the best combination of $k$ variables in a particular model (Lumley, 2009). Regsubsets() permutes all possible combinations of the remaining parameters and provides summary statistics for each permutation, including $C_p$, adjusted $R^2$, and BIC (Lumley, 2009). Equation 4.4 shows Mallow's $C_p$.

$$C_p = \frac{SSE_k}{\hat{\sigma}^2} - n + 2k \tag{4.4}$$

Equation parameters are as follows (Gilmour, 1996):

$SSE_k$: sum of the squared errors of the candidate model

$k$: number of independent variables in the candidate model

$\hat{\sigma}^2$: estimate of the mean square error from the incumbent model

$n$: number of observations

Like BIC, Mallow's $C_p$ penalizes for having extra minimal effect variables in the model, comparing all possible models with the incumbent model (Faraway, 2006). The preferred model is that with a $C_p < 2k$ (Gilmour, 1996). A $C_p$ check prior to model performance checks indicated that the model was not overfit.

Once models passed this final check, we completed diagnostics of the five assumptions to determine if any variable transformations were required (see Appendix C).

The final review was a performance test of the model using an out-of-sample data set. Validation of all MLR models was achieved with the predict() function and 2012 data to predict 2012 accessions (R Core Team, 2013). The performance of each model and summary statistics are covered in the results section.

### 3. MLR Model Results

Model performance was checked on out-of-sample 2012 data to predict 2012 accessions. The predict() function in R was used to predict accessions with a 95% prediction interval (PI). The 95% prediction interval provides an interval based on the standard error calculated from the all the $\hat{y}_i$ for each model. It was anticipated that $y_i$ would have been captured by the interval 95% of the time; however, due to random or unexplained error, the $y_i$ was captured by the PI only 91% of the time. Table 8 provides a summary view of the performance results. A full explanation of model assumptions and diagnostics is available in the Appendix C with associated plots and test results.

Table 8. Summary statistics for goodness-of-fit (adjusted $R^2$) of each regional model (national, east, and west), counts for the observations captured by the 95% prediction interval (PI), the fraction of observations in that PI, and the number of independent variables in the model. For number of observations within the 95% PI, we divided the observations within the PI by the total number of observations in that region.

| | In-Sample Adjusted $R^2$ | Out-of-Sample Adjusted $R^2$ | Number of observations ($y_i$) within 95% PI | Percent of observations ($y_i$) within 95% PI | Number of independent variables in model |
|---|---|---|---|---|---|
| National | 59% | 46% | 862/952 | 91% | 7 |
| East | 60% | 43% | 425/465 | 91% | 6 |
| West | 56% | 43% | 446/487 | 91% | 5 |

### *a. National Results*

Data used for the national model are all 952 observations (e.g., all recruiting stations in the U.S.). Because of a square root transformation to the response variable, accessions, interpretation of the regression is not very intuitive at first glance. In order to interpret the model results, especially after a transformation of the response or predictor variables, a user must understand the non-transformed effects of each coefficient coupled with some standardized $x$ value that was used in the model. Additionally, a casual observation is not enough to determine the effect each variable has in the model, as some of the coefficients can be deceiving. For example, the variables in the national model range from a minimum of 0 recruiters to a maximum of 269,773 households within AGI < $25,000. So, to understand the true effects of the coefficients and each variable in the model, Table 10 on page 39 was created using the median $x_i$ values of the independent variables to show the impact on the response.

Table 9. Final fitted station level MLR for the national region; response variable is the square root of the number of accessions.

| Independent Variable | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.937 | 0.142 | 20.743 | 0.00 |
| average number of recruiters per year | 0.430 | 0.028 | 15.192 | 0.00 |
| number of NRSs within 50 miles | 0.019 | 0.002 | 7.894 | 0.00 |
| mean unemployment rate | 0.067 | 0.015 | 4.348 | 0.00 |
| unemployment compensation | –4.535 | 0.915 | –4.958 | 0.00 |
| AGI < $25,000 | 5e–06 | 1e–06 | 4.978 | 0.00 |
| AGI > $200,000 | 3e–05 | 5e–06 | –6.025 | 0.00 |
| male veterans ages 35–54 | 2e–04 | 2e–05 | 15.411 | 0.00 |

Table 9 shows that the greatest positive effect was from recruiters. For every 1 unit of change in recruiters, we have a .43 unit of change in square root of accessions. Of important note, because the response was the square root of accessions, it would be incorrect to square .43 and claim that the result was the expected increase in number of accessions. Table 10 standardizes the $x$ values to show the decision maker the non-transformed change in the response for each variable in the model. Using Table 10, a decision maker could determine that the second most important variable affecting accessions was the number of male veterans ages 35–54, resulting in .788 non-

transformed units of positive change. If increasing the number of recruiters was not an option for a particular station, then station realignment to stack a station with ZIP codes rich in veterans ages 35–54 could prove an effective alternative.

Table 10. Contribution of each IV to accessions at the median value for each variable. For AGI, the Median $x$ value is the number of households in that station boundary that fall into the AGI category.

| Independent Variable | Coefficient | Median $x$ | Non-Transformed Effect |
|---|---|---|---|
| average number of recruiters per year | 0.430 | 2.42 | 1.041 |
| number of NRSs within 50 miles | 0.019 | 8.74 | 0.166 |
| mean unemployment rate | 0.067 | 8.96 | 0.600 |
| unemployment compensation | –4.535 | 0.09 | –0.408 |
| AGI < $25,000 | 5e–06 | 56840 | 0.284 |
| AGI > $200,000 | –3e–05 | 3792 | –0.114 |
| male veterans ages 35–54 | 2e–04 | 3941 | 0.788 |

The greatest negative effect, at first glance, appears to be unemployment compensation with a coefficient of –4.535. Furthermore, when observing the AGI > $200,000 coefficient, it would appear the effect was much less than that of unemployment compensation by a factor of 152,000. The difference between unemployment compensation and AGI > $200,000 is not as great when observing the non-transformed effects; only a factor of four.

Understanding how the coefficients change the response, the following is an explanation of each main effect. The findings from the linear models were consistent with Pinelis (2011), who modeled many of the same variables, but at the ZIP code-level. Recruiters are the primary driver of accessions. Exploration of the relationship between recruiters and accessions was an area for further research. Gibson (2009) suggested the relationship was linear up to four recruiters. Figure 3 suggests that a positive linear relationship was present up to 6 recruiters. Beyond that it was inconclusive. These findings are consistent with Gibson (2009). Gibson's findings showed a "positive association between recruiters and accessions diminishes at about four recruiters" (Gibson, 2009, p. 31).

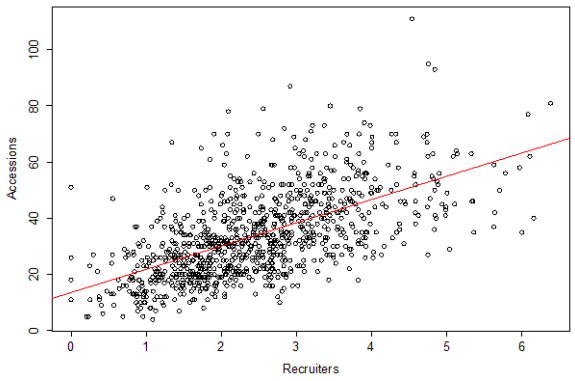


Figure 3. Indicative of a strong positive linear relationship between recruiters and accessions.

The next strongest positive effect in the model was the number of male veterans aged 35–54. Mean unemployment rate also had a strong positive influence on the number of accessions. This is consistent with Captain Jackson who identified QMA and unemployment rate as the principal socio-economic factors driving U.S. Army accessions (Jackson, 2015). One potential explanation is that as the unemployment rate increases for a given area, those individuals who lack a college degree would be more inclined to join the military.

AGI Category 1 was the number of households where the adjusted gross income was less than $25,000 per year. It had a strong positive association with accessions. In this category, the number of opportunities for youth was severely limited (Kleykamp, 2006). Kleykamp notes that, "the military may provide a source of social mobility for disadvantaged minorities during service because of the less discriminatory environment and steady employment that provides numerous benefits and compensation over other civilian equivalent jobs" (Kleykamp, 2006). The presence of mean unemployment rate and AGI < $25,000 in the national model support Kleykamp's research. One potential explanation is that the options of college and a decent paying occupation are not usually within reach of people in this category.

The number of mean number of NRSs within 50 miles of a ZIP code centroid for all ZIP codes in a given station boundary was positively correlated with the number of

accessions. One potential explanation is that as the number of recruiting stations increases, so the does the military influence and the number of recruiters.

Unemployment compensation and AGI > $200,000 were the only two variables negatively correlated with accessions in the model. Unemployment compensation was a variable that has not been previously explored in any of the studies reviewed. In fact, it has quite the opposite effect from the one expected. One potential explanation is that unemployment compensation dis-incentivizes youth, ages 17–24, to seek employment in the military, was supported by model variable selection. AGI > $200,000 was the other model variable with a negative impact on accessions. A possible explanation of the negative relationship is that families in this demographic believe in the value of a university education and have the means to pay it.

### *b. East Region Results*

Data used for the national model are all 465 observations (e.g., all recruiting stations in the east region). The east region final regression model consisted of the same IVs as the national model with the exception of two. The east region model included the total veterans variable in lieu of the male veterans ages 35–54 variable and it excluded the mean unemployment rate variable altogether. The total veterans variable had the strongest impact on the number of accessions, more so than recruiters, which was the strongest driver of accessions in the national model. Veterans, traditionally, have a strong impact in the community with military awareness and recruitment; moreover, their presence, especially those who were veterans of more popular wars (e.g., WWII) tend to have greater impact (Pinelis et al., 2011).

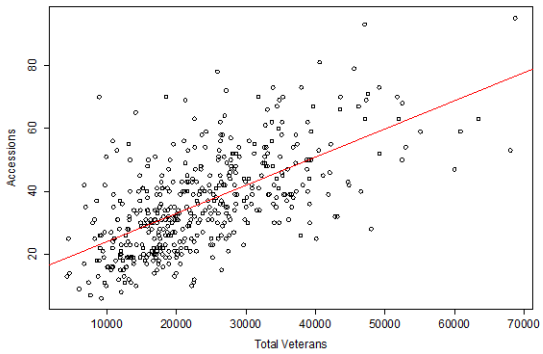


Figure 4. Accessions regressed with total veterans indicates a very strong linear relationship.

The non-transformed effects of the remainder of the variables in the model were relatively the same as those in the national model. Table 11 displays the final fitted model for the east region. Table 12 shows the non-transformed relationships between the independent variables.

Table 11. Fitted station level MLR model for the east region; response is the square root of the number of accessions.

| Parameter | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.946 | 0.174 | 22.650 | 0.00 |
| average number of recruiters per year | 0.400 | 0.044 | 8.947 | 0.00 |
| number of NRSs within 50 miles | 0.016 | 0.004 | 4.152 | 4e–05 |
| unemployment compensation | –7.725 | 1.364 | –5.664 | 0.00 |
| AGI < $25,000 | 1e–05 | 2e–06 | 5.499 | 0.00 |
| AGI > $200,000 | –5e–05 | 7e–06 | –7.020 | 0.00 |
| total veterans | 5e–05 | 5e–06 | 9.211 | 0.00 |

Table 12. Effect of the coefficient on the response using the median value from the 2011 east data set. For AGI, the Median $x$ value is the number of households in that station boundary that fall into the AGI category.

| Parameter | Coefficient | Median $x$ | Non-Transformed Effect |
|---|---|---|---|
| average number of recruiters per year | 0.400 | 2.379 | 0.952 |
| number of NRSs within 50 miles | 0.016 | 9.000 | 0.148 |
| unemployment compensation | –7.725 | 0.091 | –0.702 |
| AGI < $25,000 | 1e–05 | 622340 | 0.622 |
| AGI > $200,000 | –5e–05 | 3934 | –0.197 |
| total veterans | 5e–05 | 21544 | 1.077 |

### *c. West Region Results*

Data used for the national model are all 487 observations (e.g., all recruiting stations in the west region). The west model was the most different model out of the three MLRs. Three of its five variables are not included in east or national models. It was the only model to include university data in any of the final fitted models. Recall that universities within 50 miles were determined as the mean number of universities within 50 miles of a ZIP code centroid for all ZIP codes within a station's boundary. The effect of number of Division I universities within 50 miles of ZIP code centroid was positive. A one unit increase in division one universities resulted in a .042 unit increase in accessions. Table 13 displays the final fitted model from the west region where the response was the cubed root of the number of accessions. Table 14 displays the non-transformed effect using the median value for the independent variable.

Table 13. Fitted station level MLR model for the west region; response variable is the cubed root of the number of accessions.

| Independent Variable | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.124 | 0.04302 | 49.382 | 0.00 |
| average number of recruiters per year | 0.195 | 0.01499 | 13.033 | 0.00 |
| 2010 population | 5e–07 | 2e–07 | 3.065 | 0.00 |
| AGI > $200,000 | –1e–05 | 3e–06 | –4.165 | 0.00 |
| number of Division I universities within 50 miles of ZIP code centroid | 0.042 | 0.006 | 6.650 | 0.00 |
| male veterans ages 18–34 | 2e–04 | 2e–05 | 8.472 | 0.00 |

Table 14. Effect of the coefficient on the response using the median value from the 2011 west data set. For AGI, the Median $x$ value is the number of households in that station boundary that fall into the AGI category.

| Independent Variable | Coefficient | Median $x$ | Non-Transformed Effect |
|---|---|---|---|
| average number of recruiters per year | 0.195 | 2.463 | 0.481 |
| 2010 population | 5e–07 | 2.66e+05 | 0.144 |
| AGI > $200,000 | –1e–05 | 3361 | –0.033 |
| number of Division I universities within 50 miles of ZIP code centroid | 0.042 | 2 | 0.084 |
| male veterans ages 18–34 | 2e–04 | 1215 | 0.255 |

AGI > $200,000 had a negative impact on the number of accessions. Its non-transformed effect was weakest of all the main effects; however, it was still statistically significant with a p-value < .01. The number of recruiters had a significant impact on accessions. The effect relative to the other model variables depends on the value of the dependent variable.

From Table 14, population was the third most significant variable in the west model with a non-transformed effect of .144 units of accessions. One explanation is that areas with higher populations should yield more recruits; however, as our findings have shown, there are more key variables to predicting recruits than just the size of the pool.

## B. MODELING ZIP CODE-LEVEL ACCESSIONS

### 1. Zero-Inflated Model Selection

This section describes the rationale for use of the zero-inflated model over the traditional generalized linear model (GLM) and then provides support for choosing the zero-inflated negative binomial (ZINB) regression.

#### *a. Zero-Inflated Model*

Zero-inflated models are often used in the medical and insurance industry to model different phenomena that occur among a population where there exist an excessive number of structural zeros (Chin, 2015; Zeileis, 2008). Structural zeros are those observations where the response has zero probability mass, given the independent variables in the model (Chin, 2015). Excessive zeros can present a problem for data analysis (Chin, 2015). The zero-inflated model combines two components jointly into one simultaneous model using a binomial and count model (Chin, 2015). The logistic model determines whether or not an observation will produce structural zeros with a probability $\pi = f_{zero}(0; z, \gamma)$ (Zeileis, 2008). The second estimates the parameters of the model for the non-structural zeros. The model for the non-structural zeros includes random zeros. The mean of the probability of observing a random or "non-structural" zero is $1-\pi$ and can be found in Chin (2015) and Zeileis (2008).

The full zero-inflated density with the zero-inflated component and the count component is:

$$f_{ZEROINFL}(y;x,z,\beta,\gamma) = f_{ZERO}(0;z,\gamma) * I_{\{0\}}(y) + (1 - f_{ZERO}(0;z,\gamma)) * f_{COUNT}(y;x,\beta) \quad (4.5)$$

Parameters are defined as follows:

$y$: response variable

$x$: vector of independent variables in the count component of the model

$z$: vector of independent variables in the zero component of the model

$\beta$: vector of independent variable's coefficients in the count component of the model

$\gamma$: the vector of independent variable's coefficients in the zero component of the model

The two components of the model are as follows:

$I_{\{0\}}(y)$: structural zero component = 0, if $y \neq 0$

$I_{\{0\}}(y)$: structural zero component = 1, if $y = 0$

$f_{COUNT}(y;x,\beta)$: count distribution component

Now, in direct application of the zero-inflated model to our research, the $z$ variables are those variables, which increase the probability of observing a structural zero, and the $x$ variables are those variables that increase the probability of observing an accession and measuring that count. Selection of variables $x$ and $z$ is covered later in the chapter.

### *b. Zero-Inflated Negative Binomial Regression*

A frequency distribution of the response is one way to determine if a zero-inflated model is appropriate for the data (Flynn, 2009, p. 176). Figure 5 shows the excess distribution of zeros. On average, 64% of the ZIP codes contained zero accessions for the year.

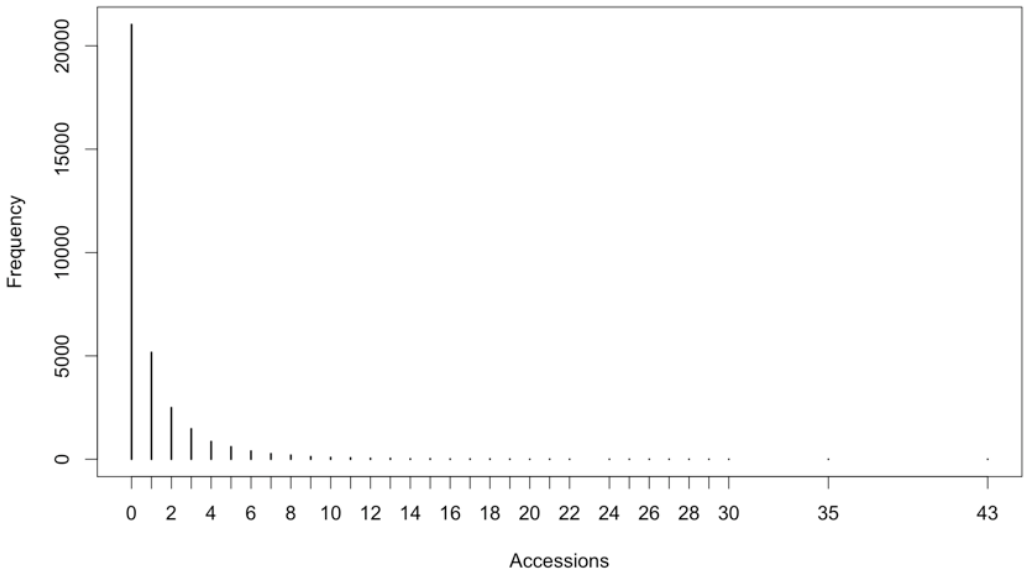


Figure 5. Distribution of national level accessions at the ZIP code-level indicating an excess count of zeros.

The response contains 64% zeros, indicating the need for a zero-inflated model. The next decision was the distribution for the number of accessions that have the possibility of being positive. The count component of the zero-inflated model is often specified as having a standard Poisson, geometric, or negative binomial GLM to predict the count from that observation. The dispersion (under or over-dispersion) of the data suggests the distribution to use with the zero-inflated model. Since the variance > mean for the response (see Table 16), a ZINB was indicated (Rodriguez, 2013). During the initial model fitting, the zero-inflated Poisson regression models did not converge. The zero-inflated negative binomial models did converge. Tables 15 and 16 provide the mean and variance of accessions for all three data sets, with and without zeros, respectively.

Table 15. Summary statistics of the response variable, number of accessions, showing the variance is greater than the mean indicating over-dispersion. Table represents all zip code level accessions data for the respective region.

| Metric | National | East | West |
|---|---|---|---|
| Mean | .99 | .98 | 1.00 |
| Variance | 4.32 | 3.76 | 4.93 |
| Zero Observations | 21,034 | 10,592 | 10,442 |

Table 16. Summary statistics of the response variable, number of accessions, with zero observations removed, showing the variance is greater than the mean indicating over-dispersion.

| Metric | National | East | West |
|---|---|---|---|
| Mean | 2.73 | 2.52 | 3.02 |
| Variance | 7.15 | 5.80 | 8.75 |

When the frequency distribution of zeros is less obvious, another method of determining if the zero-inflated model is appropriate is the Vuong test in R (Gibson, 2009; R: Jackman, 2015). The Vuong test compares the zero-inflated model to the GLM with the same independent variables using a non-nested hypothesis test (Jackman, 2015). Despite the high percentage of zeros, we confirmed that a zero-inflated model was appropriate using the Vuong test (Jackman, 2015). Results for the Vuong test can be observed in Table 17.

Table 17. Vuong Non-Nested Hypothesis Test-Statistic. M1 is a ZINB model fitted national ZIP code-level accessions and M2 is a Poisson GLM fitted to the same data set.

| | Vuong z-statistic | p-value |
|---|---|---|
| Raw | 32.3, M1 > M2 | <2e–16 |
| AIC-corrected | 32.2, M1 > M2 | <2e–16 |
| BIC-corrected | 32.0, M1 > M2 | <2e–16 |

The null hypothesis of the test is that M1 and M2 are indistinguishable, but given a p-value < .05, it is clear that we would reject the null. The Vuong statistic is distributed under the null hypothesis as a N(0,1) (Jackman, 2015). A Vuong statistic > 1.96 indicates that M1 is a superior model to M2 (Ismail, 2013; Jackman, 2015).

### 2. The Zero-Inflated Negative Binomial Regression

Now that we have identified the basis for selection of the zero-inflated model and selection of the distribution for the zero-inflated model, we will describe the ZINB model. The negative binomial model is observed as a "generalization of the Poisson where the [mixing] parameter $\gamma$ is gamma distributed" and the logarithm of the mean of

the negative binomial distribution by a linear combination of independent predictor variables (Faraway, 2006, p. 71). The negative binomial model, unlike the Poisson model, does not assume that mean and variance are equal; therefore, the negative binomial corrects for over-dispersed data. The ZINB model predicts the log of the number of accessions, so the interpretation of the coefficients follows: for every single unit change in $x_i$, the modeler can expect the log of the mean of $y_i$ to change by the value of the independent variable coefficient (Piza, 2012).

## 3. Model Development and Variable Selection

This chapter continues with a brief description of the variable selection techniques used for the ZINB model and an in-depth explanation of the model performance results on the test data and the inferences made from the results for the national level model only. Tables with east and west region model results can be found in the appendix.

We leveraged our knowledge from our variable selection process from our station level MLR model to inform the subset of IVs we started with for the zero-inflated variable selection process. Each zero-inflated modeling process began with the same 17 variables (see Figure 6 and Appendix A for variable descriptions). The same 17 were used for both the logistic model and negative binomial model. Figure 6 shows the correlations between the independent variables of the subset of variables used for the ZIP code-level models.

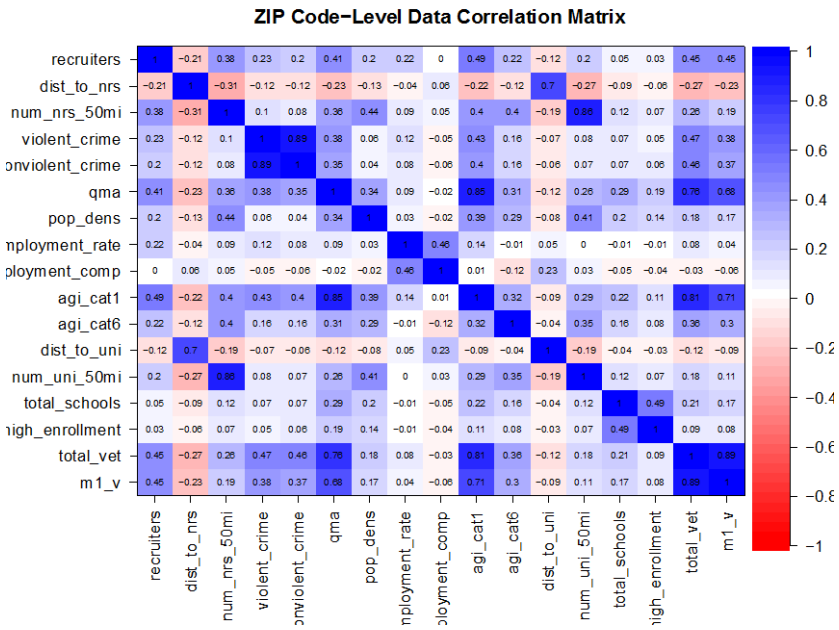


Figure 6. The correlation matrix exhibits a high degree of correlation among the subset of covariates used for the initial variable selection process.

A backward stepwise regression was used with the be.zeroinfl() function in the mpath package in R (Zhu Wang, 2015). The be.zeroinfl() takes, as its main argument, a zeroinfl() object and then according to the significance level criterion eliminates statistically insignificant variables (Zhu Wang, 2015). Statistical significance for ZIP code-level models, unless otherwise stated, is a p-value < .01. This process was used to fit three models using different subsets of the data, national, east, and west. Prior to evaluating performance statistics described below, models were evaluated based on the minimum Bayesian Information Criterion (BIC).

## 4. Zero Inflated Model Results

Model performance statistics were calculated through a series of evaluations after predicting 2012 accessions. A true negative is the ability of the logistic model or negative binomial model to correctly predict the zero accessions ZIP codes that affect station closures. We used three other similar metrics, including true positives, false positives, and false negatives. A true positive is the negative binomial model's ability to correctly identify the precise non-zero number of accessions in a ZIP code. A false positive occurs

when the negative binomial model incorrectly identifies a ZIP code as having greater than zero accessions, when the actual number of accessions was zero. Lastly, a false negative is the instance when the logistic model incorrectly predicts zero. This was important because in data from a zero-inflated model (i.e., 64% zeros), the model's predictive capability can be deceptively accurate if it just defaulted to zero most of the time, so it was important to track how often it labels a ZIP code zero, when there were actually accessions.

The other two metrics are mean absolute deviation (MAD) for the model and for the negative binomial portion of the model. The formula for mean absolute deviation is in equation 4.6, where $n$ is the number of observations. It is a precise measurement of the average inaccuracy of the model, measured in accessions.

$$MAD_{MODEL} = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{n} \tag{4.6}$$

The following section contains the performance results for the national ZINB model, using 2011 data to fit and 2012 to validate. Performance results for the east and west models are contained in Appendix D.

### *a. ZINB National Results*

Using a level of significance, p-value < .01, the fitted national ZINB model consisted of 8 independent variables for the logistic component and 11 independent variables for the negative binomial component of the model. Table 18 contains the results from the logistic model and Table 19 contains the results from the negative binomial model. The MAD, false positive rate, and other metrics are contained in Table 22 in Chapter V.

In Table 18, variables with positive estimated coefficients, increase the probability of observing a zero for that ZIP code. As expected, households with an AGI greater than $200,000, the number of universities within 50 miles, and the total number of universities all increase the probability of observing a zero in a particular ZIP code.

Total schools had the greatest positive coefficient of .636. This was consistent with our findings at the station level.

Unemployment compensation decreased the probability of observing a zero for a particular ZIP code. This was not consistent with our findings at the station level. AGI greater than $200,000, the number of universities within 50 miles, and the total number of universities are present in both, the negative binomial and logistic models, increasing the chance of observing a structural zero for a particular ZIP code. Unemployment compensation had the same coefficient sign for both models. Because the unemployment compensation coefficient is so small in the negative binomial model we would expect little practical impact from this variable in the count model.

As expected, recruiter numbers, nonviolent crime, QMA, and total veterans had strong negative effects in the logistic model. This is consistent with our findings at the station level.

Table 18. Summary view of the logistic component of the national ZINB model

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 2.333 | 0.082 | 28.579 | 0.000 |
| average number of recruiters per year | –0.016 | 0.006 | –2.736 | 0.006 |
| reports of violent crime | –0.013 | 0.002 | –5.597 | 0.000 |
| QMA | –0.005 | 3e–04 | –17.175 | 0.000 |
| unemployment compensation | –0.021 | 0.006 | –3.470 | 0.001 |
| AGI > $200,000 | 0.001 | 2e–04 | 5.314 | 0.000 |
| number of Division I universities within 50 miles of the ZIP code centroid | 0.031 | 0.010 | 3.038 | 0.002 |
| total number of universities in the ZIP code | 0.636 | 0.241 | 2.639 | 0.008 |
| total number of veterans | –0.003 | 2e–04 | –13.495 | 0.000 |

Considering Table 19 and the count model, recruiters, the number of NRSs within 50 miles, violent crime, mean unemployment rate, households with an AGI less than $25,000, and total veterans all had a positive statistically significant association with accessions. Violent crime was unexpected, as this was not statistically significant in Gibson (2009) or Pinelis et al. (2011). One explanation of its significance is that in areas

with high crime rates and poverty, these youth see the military as a way out (Kleykamp, 2006). Mean unemployment rate, consistent with Jackson (2015), increased the predicted number of accessions.

Distance to the nearest NRS had a negative impact on the number of accessions. This was consistent across all models and with Pinelis et al. (2011). Households with an AGI greater than $200,000, the number of universities within 50 miles, and the total number of universities all decreased the predicted number of accessions. This is consistent with the logistic model where these variables increased the probability of classifying a particular ZIP code as a structural zero. Based on this finding, NRC would likely have difficulty recruiting in affluent areas with a high density of universities.

Table 19. Summary view of the negative binomial component of the national ZINB model

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | –0.414 | 0.038 | –10.914 | 0.000 |
| average number of recruiters per year | 0.014 | 0.001 | 17.957 | 0.000 |
| distance to NRS | –0.007 | 0.001 | –9.707 | 0.000 |
| number of NRSs within 50 miles | 0.010 | 0.001 | 8.457 | 0.000 |
| reports of violent crime | 0.001 | 2e–04 | 4.910 | 0.000 |
| mean unemployment rate | 0.014 | 0.004 | 3.583 | 0.000 |
| unemployment compensation | –0.007 | 0.003 | –2.853 | 0.004 |
| AGI < $25,000 | 5e–05 | 2e–06 | 26.979 | 0.000 |
| AGI > $200,000 | –1e–04 | 2e–05 | –5.556 | 0.000 |
| number of Division I universities within 50 miles of the ZIP code centroid | –0.019 | 0.003 | –6.557 | 0.000 |
| total number of universities in the ZIP code | –0.057 | 0.017 | –3.361 | 0.001 |
| total number of veterans | 4e–04 | 1e–05 | 29.428 | 0.000 |
| Log(theta) | 1.517 | 0.009 | 174.976 | 0.000 |

The MAD for the national model was the highest at .73, meaning that our average model prediction was off on average by .73 annual accessions at the ZIP code-level. This can be compared to that of Pinelis et al. model, which estimated a MAD of .943; however, it should be noted that different data were used for the regression models. One area of further research recommended by Pinelis was “correctly predicting nonproductive

ZIP codes" (Pinelis, 2011). NRC has placed a premium on the ability to accurately predict a true negative (D. Ammons-Moreno, personal communication, August 20, 2015). A true negative is the percentage at which a ZINB model correctly predicts zeros. Our models performed well with respect to this metric. The national model had a true negative rate of .81, but the west model had a true positive rate of .86. Table 22 provides all statistics from the performance tests of all three ZINB models. The log($\theta$) line, in Table 19, describes the mixing parameter of the negative binomial component model.

For further analysis of the ZINB models, figures 7 and 8 are provided. The boxplot in Figure 7 graphically represents the difference between the actual number of accessions and the predicted number of accessions for observed number of accessions from 0 to 18. We chose 18 because fewer than 4% of the actual accessions came from ZIP codes that produced more than 18 accessions in 2012. Figure 7 suggests the following:

1. There is a minimal slight positive trend from zero actual accessions out to 18 indicating that for ZIP codes with higher accessions, we can expect a greater error between the actual and predicted. Considering more than 96% of the accessions in 2012 came from ZIP codes where the mean residuals are inside –1 and +5, the model indicates a high degree of precision when actual accessions are less than 18.

2. When actual accessions are between two and six, 50% of the residuals are consistent between +/– 1.5 of the median. With observations that have less than three actual accessions, we see only a slight deviation from the median. For accessions six and greater, the middle 50% gets wider, but is still within +/– 6 of the median.

3. For the range of actual accessions from 0 to 18, the predicted accessions are greater than actual accessions. This trend tends to decrease as the actual number of accessions increases. The reason for this is unknown.

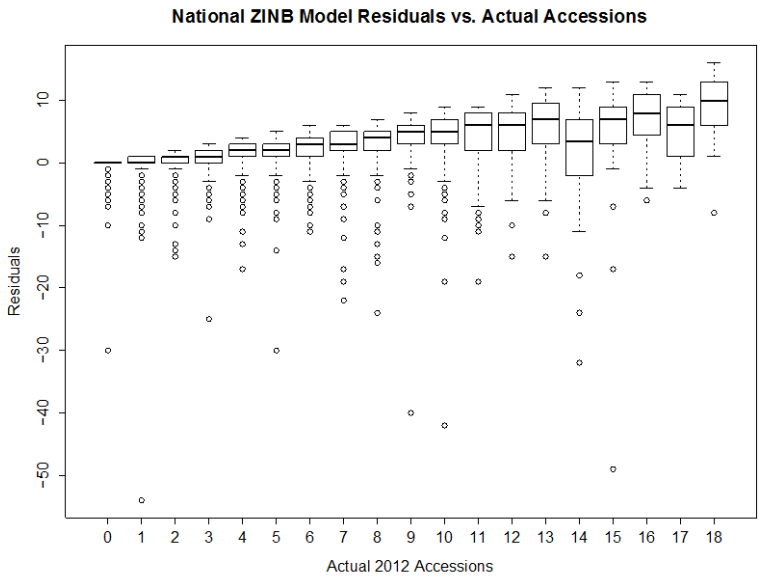


Figure 7. Plot of the residuals versus the actual accessions. Residuals are actual accessions minus predicted accessions.

In 2012, considering ZIP codes where the actual accessions > 10, 1% of these ZIP codes produced 14% of the accessions. We recommend NRC devote more attention to these ZIP codes that are producing 10 and fewer, because it is from these ZIP codes where 86% of the Navy's newest accessions are hailing. Figure 8 shows a micro boxplot of Figure 7 taking only these observations into account.

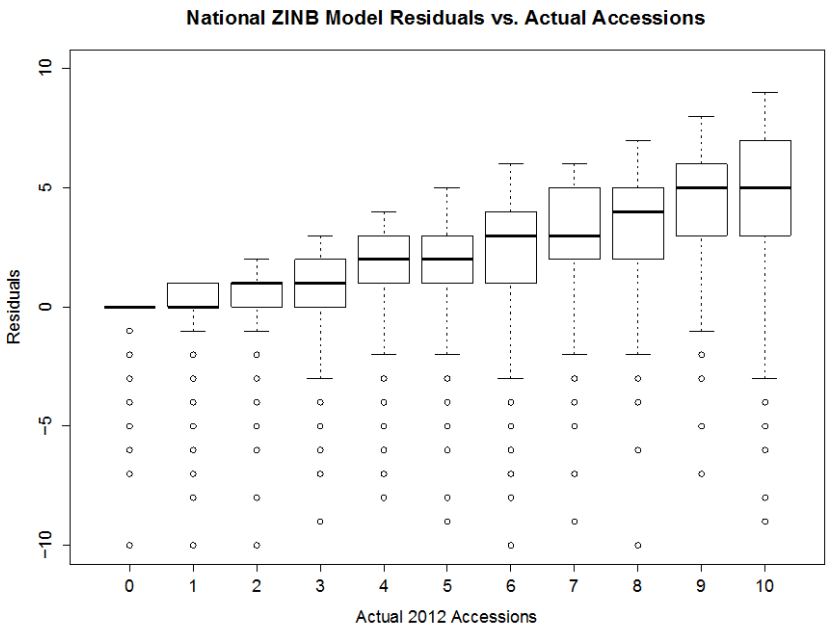


Figure 8. Micro boxplot showing observations with < 11 accessions and 0 +/– 10 residuals. This ZINB national model's predictive capability is greatest in the ZIP codes where Navy recruits 86% of its accessions, which is why NRC should implement this model for predicting zeros and lower producing ZIP codes.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. SUMMARY AND CONCLUSIONS

## A. RECOMMENDATIONS

### 1. Station Level (MLR Models)

If incorporated, these models will aid district leadership with manning stations, determine stations trends, and help to set realistic annual goals. Overall, the number of recruiters was the strongest indicator of the number of accessions a station could expect; so increasing recruiters at a station would be the most viable option to increasing accessions. This was only to a certain point, as further work is required to determine the point of saturation or where the production levels plateau with the increase in recruiters.

The next best option to increase accessions at the station level was to realign ZIP code boundaries to include ZIP codes with large populations of veterans in the east region, as total veteran population was the strongest indicator of accession in this model.

Table 20. Consolidated summary of all MLR station level models.

| Model | National | East | West |
|---|---|---|---|
| Adj $R^2$ | .59 | .60 | .56 |
| 95% PI Validation Summary | 862/952 | 862/952 | 862/952 |
| Model Assumptions | Meets all model assumptions | Meets all model assumptions | Meets all model assumptions |
| IV with greatest positive impact on accessions | average number of recruiters per year | total veterans | average number of recruiters per year |
| IV with greatest negative impact on accessions | unemployment compensation | unemployment compensation | AGI > $200,000 |
| Number of IVs | 7 | 6 | 5 |

All models included AGI > $200,000. This variable strongly indicated that youth in these areas are less propensed to enlist. Avoiding these ZIP codes would increase the number of accessions and minimize time wasted by recruiters. The IV, AGI < $25,000, was present in the national and east models. This predictor variable was indicative of

high accessions, so to recruit in these ZIP codes and re-zone these ZIP codes to poorer performing stations could prove to shift the balance to under-producing stations.

### 2. ZIP Code-Level (ZINB Models)

If incorporated, these models will aid station leadership in identifying the high schools, by ZIP code, that will produce the greatest number of accessions. Furthermore, the ZINB model will help in ZIP code re-alignment of stations. Many of the independent variable findings from the station level models reflect those of the ZIP code-level models. The ZIP code-level model solves NRC's problem of level of analysis and provides accurate predictions of both structural zero ZIP codes and counts for ZIP codes. The ZINB models and MLR models represented many of the same variables for positive and negative effects on accessions, including recruiters, AGI Category 1 and 6, and total veterans. NRC should give due consideration to these variables when considering locations for opening recruiting stations and establishing annual station goals.

Table 21. Consolidated summary of all MLR station level models.

| Model | National | East | West |
|---|---|---|---|
| BIC | 61776 | 33765 | 27837 |
| Model MAD | 0.744 | 0.732 | 0.692 |
| Model Assumptions | Meets all model assumptions | Meets all model assumptions | Meets all model assumptions |
| IV with greatest positive impact per unit change on count accessions | average number of recruiters per year and mean unemployment rate | average number of recruiters per year | mean unemployment rate |
| IV with greatest negative impact per unit change on count of accessions | total number of universities in ZIP code | total number of universities in ZIP code | unemployment compensation |
| Best predictor of structural zeros | total number of universities in ZIP code | total number of universities in ZIP code | AGI > $200,000 |

Exactly half of the accessions in the U.S. for 2012 came from the 30,597 ZIP codes that produced fewer than five accessions annually (18,458/37,032), where 37,032 are the total number of accessions from the year 2012; it is our recommendation that NRC focus its recruiting efforts on the smaller ZIP codes where the influx of recruits is

not as steady as the larger ZIP codes. The ZINB model is highly accurate in these ZIP codes as observed in Figure 8. Future research, similar to that of Jackson (2015) could explore metropolitan and non-metropolitan areas. Assuming areas that produce five or fewer accessions in a year are non-metropolitan, NRC could implement our ZINB model at that level where it performs best.

Table 22. Summary view of the performance metrics from all three models validated on their respective region data sets. See explanation of table entries in Chapter IV.

| Metric | National ZINB | East ZINB | West ZINB |
|---|---|---|---|
| Model MAD | 0.73 | 0.73 | 0.69 |
| Count MAD | 1.65 | 1.47 | 1.78 |
| True Positive | 0.279 | 0.297 | 0.254 |
| True Negative | 0.808 | 0.761 | 0.859 |
| False Positive | 0.192 | 0.239 | 0.141 |
| False Negative | 0.177 | 0.165 | 0.192 |
| BIC | 61776 | 33765 | 27837 |
| NLL | 30774 | 16794 | 13836 |
| Theta | 4.558 | 5.732 | 4.384 |
| Number of predicted zeros | 8899 | 9045 | 9805 |
| % of zeros | 0.570 | 0.522 | 0.626 |
| Number of binomial IVs | 8 | 7 | 4 |
| Number of count IVs | 11 | 9 | 10 |

## B. FUTURE WORK

During this course of this thesis, challenges and discoveries were encountered, some of which, because of a lack of time and resources, we were unable to pursue to the end. While this research quantitatively answered many of the questions regarding socio-economic factors and recruiting, improved upon previous ZIP code-level models, and created a decision making tool for individual stations, further work is warranted. Below are three recommended avenues for future study in this field.

### 1. Declining Health of Today's Youth

Chronic health problems, specifically Type 2 Diabetes, have been a growing concern in this country. During the course of our literature review, it was noted in a number of studies and articles on recruiting for military service, that obesity was a growing concern for the future of our military force. With an increase in childhood obesity rates and decrease in the level of inactivity, Type 2 Diabetes rates among adolescents has seen staggering climbs over the past two decades. According to a 2007 study in Diabetes Journal, "age-specific increases in annual hospitalizations for diabetes occurred primarily among individuals aged 20–24 years (152.6 of 100,000 in 1993 and 222.2 of 100,000 in 2004)" (Lee, 2007). With the increase in the proportions of obese adolescents and those with Type 2 diabetes, a disease disqualifying for military service, we assert that the health effects of today's youth could affect recruiting efforts in certain geographic regions with growing rates of obesity. Health was one of the original data categories we hoped to model; however, due to the lack of data available from the sources where we searched, it could not be included.

### 2. Additional Data

Nielsen's PRIZM Market Segmentation Data provides ZIP code-level data on 66 segments that were ranked according to socio-economic status. It takes into account many of the data that we used to build our models, including income and education, among others. These data categories identify subgroups of the population for corporations to help them identify the most prolific markets for their particular product. This was precisely what NRC was attempting to do—find a particular market for their product, military service. By incorporating these data into similar Army models, one previous study by Captain Liam Marmion (2015), was able to increase his explanation of error in linear models at the Company level by 10%–15%. These data were not available at the time of our data collection and model development phases and was noted in the constraints section of this study.

# APPENDIX A: META DATA

## A. ZIP CODE-LEVEL META-DATA

Appendix A provides meta-data and explains every variable. Because the data were aggregated at two separate levels, the explanations differ for some of the variables.

Table 23. Meta-Data for ZIP code-level data

| Variable | Description |
|---|---|
| year | year of the observation; 2011–2013 |
| rsid | station ID by which the data was aggregated from the ZIP code-level; 952 unique |
| nrd | recruiting district; 13 East and 13 West |
| region | east and west regions combine to make up the NRC AOR |
| accessions | the number of signed contracts from this ZIP code and year |
| num_nrs_50mi | number of NRSs inside 50 mile radius of ZIP code centroid |
| recruiter_to_qma | ratio of the number of recruiters to the number of QMA in that ZIP code |
| recruiters | the average number of recruiters assigned to ZIP code |
| dist_to_nrs | distance from ZIP code centroid to nearest NRS |
| violent_crime | weighted proportions of violent crimes based on county crime counts and ZIP code populations |
| nonviolent_crime | weighted proportions of nonviolent crimes based on county crime counts and ZIP code populations |
| qma | quality of military available aged 17–24 |
| pop_2010 | 2010 census population data |
| pop_dens | number of persons per square mile |
| mean_unemployment_rate | average of the monthly unemployment rates for each year |
| std_dv_unemployment | standard deviation of monthly rates to give us an indication of the ebb and flow<br>of the economy |
| poverty | rate of people living below the poverty line |
| avg_agi | average adjusted gross income per return divided by the total number of returns<br>(sum AGIs/total # returns)(x$1,000) |
| unemployment_comp | ratio of the number of the number of returns with unemployment compensation<br>to total number of returns |
| per_returns_taxable | ratio of the number of returns with taxable income divided by the total number<br>of returns (# returns with taxable income/total # returns) |
| pensions_annuities_agi | ratio of total amount from returns with taxable pensions and annuities to total<br>amount from returns with taxable income |

| Variable | Description |
|---|---|
| agi_cat1 | $1 < agi < $25000 |
| agi_cat2 | $25000 <= agi < $50000 |
| agi_cat3 | $50000 <= agi < $75000 |
| agi_cat4 | $75000 <= agi < $100000 |
| agi_cat5 | $100000 <= agi < $200000 |
| agi_cat6 | agi >= $200000 |
| six_fig | this was the total number of returns with greater than a six figure AGI |
| dist_to_uni | distance to the nearest Division I university from ZIP code centroid |
| num_uni_50mi | total number of Division I universities within 50 nm of ZIP code centroid |
| total_schools | total number of 4-year, public and private degree granting, not-for-profit universities in ZIP code |
| high_enrollment | considering all universities in the ZIP code, this was the highest enrollment school |
| u_pop1 | number of universities within ZIP code that have fewer than 1000 students |
| u_pop2 | number of universities within ZIP code that have 1000–4999 students |
| u_pop3 | number of universities within ZIP code that have 5000–9999 students |
| u_pop4 | number of universities within ZIP code that have 10000–19999 students |
| u_pop5 | number of universities within ZIP code that have greater than 20000 students |
| price_in_state | total price for in-state students living on campus |
| price_out_state | total price for out-of-state students living on campus |
| total_vet | total weighted veteran population based on 2011–2013 (3) year ACS county data |
| tot_m_vet | total number of male veterans |
| m1_v | 18–34 all male veterans |
| m2_v | 35–54 all male veterans |
| m3_v | 55–64 all male veterans |
| m4_v | 65–74 all male veterans |
| m5_v | 75+ all male veterans |
| tot_f_vet | total number of female veterans |
| f1_v | 18–34 all female veterans |
| f2_v | 35–54 all female veterans |
| f3_v | 55–64 all female veterans |
| f4_v | 65–74 all female veterans |
| f5_v | 75+ all female veterans |

Table 24. Meta-data for station level data

| Variable | Description |
|---|---|
| year | year of the observation; 2011–2013 |
| rsid | station ID by which the data was aggregated from the ZIP code-level; 952 unique |
| nrd | recruiting district; 13 East and 13 West |
| region | east and west regions combine to make up the NRC AOR |
| accessions | the number of signed contracts from observed station and year |
| num_nrs_50mi | mean number of NRSs inside 50 mile radius of every ZIP code centroid within the station |
| recruiter_to_QMA | ratio of the number of recruiters to the number of QMA in that station area |
| station_area | station land area covered |
| recruiters | sum of the average number of recruiters assigned to all ZIP codes within that station |
| dist_to_nrs | median of all the minimum ZIP code centroid to NRS distances in that station |
| violent_crime | summed weighted proportions of violent crimes by all ZIP codes in the station |
| nonviolent_crime | summed weighted proportions of nonviolent crimes by all ZIP codes in the station |
| qma | quality of military available aged 17–24 |
| pop_2010 | 2010 census population data |
| pop_dens | median population density of all ZIP codes within the station |
| mean_unemployment_rate | average of the monthly unemployment rates for each year |
| std_dv_unemployment | mean of all the standard deviations for the ZIP codes in the station |
| poverty | rate of people living below the poverty line |
| avg_agi | median of all the average AGIs from all ZIP codes in the station |
| unemployment_comp | ratio of the number of the number of returns with unemployment compensation<br>to total number of returns |
| per_returns_taxable | ratio of the number of returns with taxable income divided by the total number<br>of returns (# returns with taxable income/total # returns) |
| pensions_annuities_agi | ratio of total amount from returns with taxable pensions and annuities to total<br>amount from returns with taxable income |
| agi_cat1 | $1 < agi < $25000 |
| agi_cat2 | $25000 <= agi < $50000 |
| agi_cat3 | $50000 <= agi < $75000 |
| agi_cat4 | $75000 <= agi < $100000 |
| agi_cat5 | $100000 <= agi < $200000 |
| agi_cat6 | agi >= $200000 |
| six_fig | this was the total number of returns with greater than a six figure AGI |
| dist_to_uni | median of all the minimum ZIP code centroid to Division I university |

| Variable | Description |
|---|---|
| | distances in that station |
| num_uni_50mi | total number of Division I universities within 50 nm of ZIP code centroid |
| total_schools | total number of 4-year, public and private degree granting, not-for-profit universities |
| high_enrollment | considering all universities in the station, this was the population of the highest enrollment school |
| u_pop1 | number of universities within station that have fewer than 1000 students |
| u_pop2 | number of universities within station that have 1000–4999 students |
| u_pop3 | number of universities within station that have 5000–9999 students |
| u_pop4 | number of universities within station that have 10000–19999 students |
| u_pop5 | number of universities within station that have greater than 20000 students |
| total_vet | total weighted veteran population based on 2011–2013 (3) year ACS county data |
| tot_m_vet | total number of male veterans |
| m1_v | 18–34 all male veterans |
| m2_v | 35–54 all male veterans |
| m3_v | 55–64 all male veterans |
| m4_v | 65–74 all male veterans |
| m5_v | 75+ all male veterans |
| tot_f_vet | total number of female veterans |
| f1_v | 18–34 all female veterans |
| f2_v | 35–54 all female veterans |
| f3_v | 55–64 all female veterans |
| f4_v | 65–74 all female veterans |
| f5_v | 75+ all female veterans |

# APPENDIX B: DATA CLEANING

Appendix B provides the reader with equations used in the development and transformation of the raw data into the final format used in the model. Also provided are footnotes and other amplifying information for the FBI UCR data.

$$D = (((180 * 60)/\pi) * (\arcsin(\sin(y_1) * \sin(y_2) + \cos(y_1) * \cos(y_2) * \cos(x_1 - x_2))) \quad \text{(B.1)}$$

where y and x are latitudes and longitudes, respectively, of the NRS or Division I university and ZIP code centroid.

Table 25. Footnotes from the 2011 UCR data set from (Federal Bureau of Investigation, 2015).

| State and County | Note |
|---|---|
| All states and counties | [1] If a blank was presented in the arson column, it indicates that the FBI did not receive 12 complete months of arson data for that agency. |
| All AL, MT; Allegany, MD; Oakland, CA | [2] Because of changes in the state/local agency's reporting practices, figures were not comparable to previous years' data. |
| Graham, AZ; Rock, MN; Greenwood, SC | [3] The FBI determined that the agency's data were over-reported. Consequently, affected data were not included in this table. |
| Tulare, CA | [4] The Tulare County Highway Patrol collects the motor vehicle thefts for this county. These data can be found in Table 11 (see FBI UCR). |
| Pierce, GA | [5] The FBI determined that the agency's data were underreported. Consequently, those data were not included in this table. |
| Harrison, IN | [6] The FBI determined that the agency did not follow national Uniform Crime Reporting (UCR) Program guidelines for reporting an offense. Consequently, this figure was not included in this table. |
| All MN counties | [7] The data collection methodology for the offense of forcible rape used by the Minnesota state UCR Program does not comply with national UCR Program guidelines. Consequently, its figures for forcible rape and violent crime (of which forcible rape was a part) were not published in this table. |

Table 26. Footnotes from the 2012 UCR data set from (Federal Bureau of Investigation, 2015).

| State and County | Note |
|---|---|
| All states and counties | [1] The FBI does not publish arson data unless it receives data from either the agency or the state for all 12 months of the calendar year. |
| Pinal, AZ; 11 counties in LA; Valencia, NM; | [2] The FBI determined that the agency's data were under-reported. Consequently, those data were not included in this table. |
| Graham, AZ; El Paso, CO; Cibola, NM; 3 counties in TX | [3] The FBI determined that the agency's data were over-reported. Consequently, affected data were not included in this table. |
| Tulare, CA | [4] The Tulare County Highway Patrol collects the motor vehicle thefts for this county. These data can be found in Table 11 (see FBI UCR). |
| Douglas, CO; Hamilton, OH | [5] Because of changes in the state/local agency's reporting practices, figures were not comparable to previous years' data. |
| All MN counties | [6] The data collection methodology for the offense of forcible rape used by the Minnesota state Uniform Crime Reporting (UCR) Program does not comply with national UCR Program guidelines. Consequently, its figures for forcible rape and violent crime (of which forcible rape was a part) were not published in this table. |
| Rock, MN | [7] The FBI determined that the agency did not follow national UCR Program guidelines for reporting an offense. Consequently, this figure was not included in this table. |

Table 27. Footnotes from the 2013 UCR data set from (Federal Bureau of Investigation, 2015).

| State and County | Note |
|---|---|
| All states and counties | [1] The figures shown in this column for the offense of rape were reported using the revised Uniform Crime Reporting (UCR) definition of rape. See Data Declaration for further explanation. |
| All states and counties | [2] The figures shown in this column for the offense of rape were reported using the legacy UCR definition of rape. See Data Declaration for further explanation. |
| All states and counties | [3] The FBI does not publish arson data unless it receives data from either the agency or the state for all 12 months of the calendar year. |
| Pinal, AZ; Tulare, CA; Nye, NV | [4] The FBI determined that the agency's data were under-reported. Consequently, those data were not included in this table. |
| Glynn, GA | [5] The FBI determined that the agency's data were over-reported. Consequently, those data were not included in this table. |
| Shelby, IN; 4 counties in MS | [6] This agency began the year submitting rape data classified according to the legacy UCR definition. However, at some point during the calendar year, the agency modified its reporting methods and began classifying and submitting rape offenses according to the revised UCR definition of rape. See Data Declaration for further explanation. |
| Franklin, ME; Franklin, PA | [7] Because of changes in the state/local agency's reporting practices, figures were not comparable to previous years' data. |

# APPENDIX C: MLR MODEL

Appendix C provides the reader with the justification for variable transformation, validation of the linear model assumptions, and other information germane to each model's fit.

## A. NATIONAL MODEL

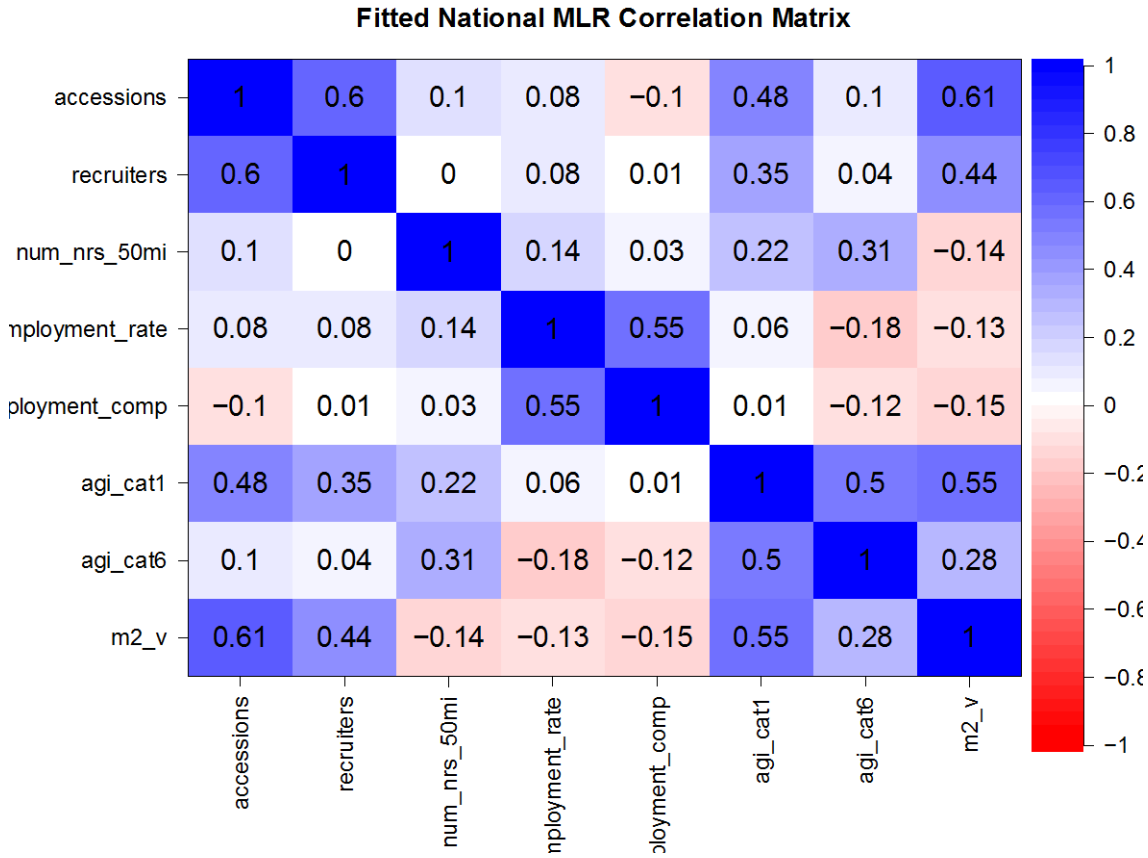


Figure 9. Correlation matrix showing the degree of correlation between the final subset of predictor variables and the response variable.

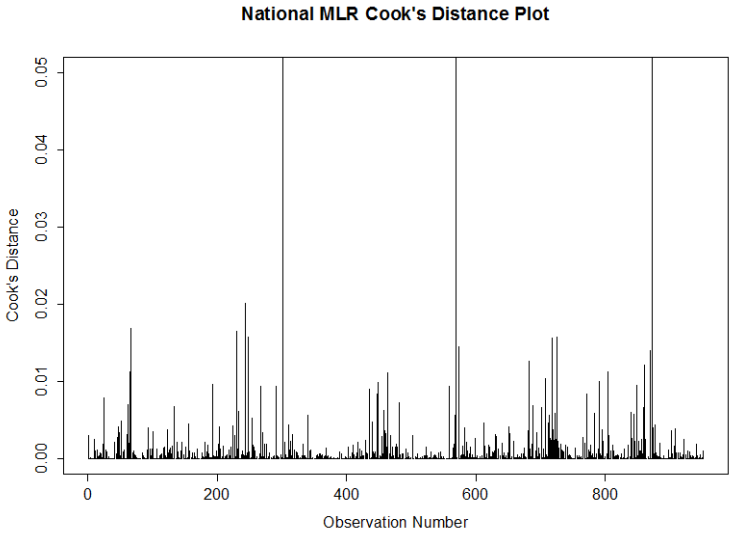


Figure 10. The fitted national model contained three outliers, as defined by Cook's distance > .05. Outliers were removed and model fit improved.

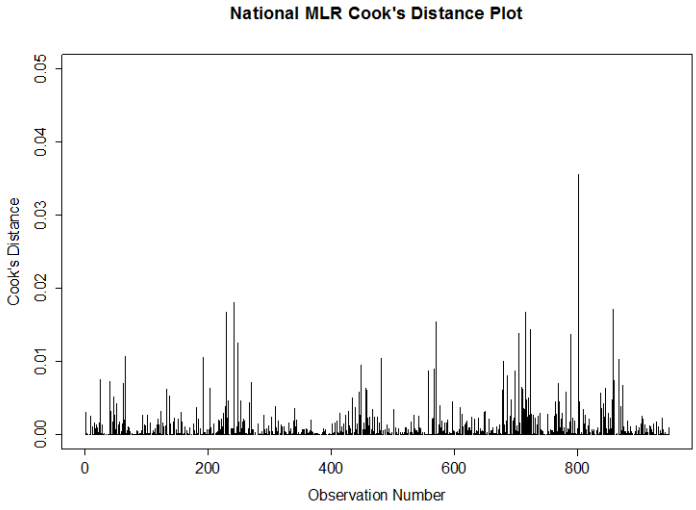


Figure 11. Plot showing the Cook's distance of all observations in final fitted model without outliers.

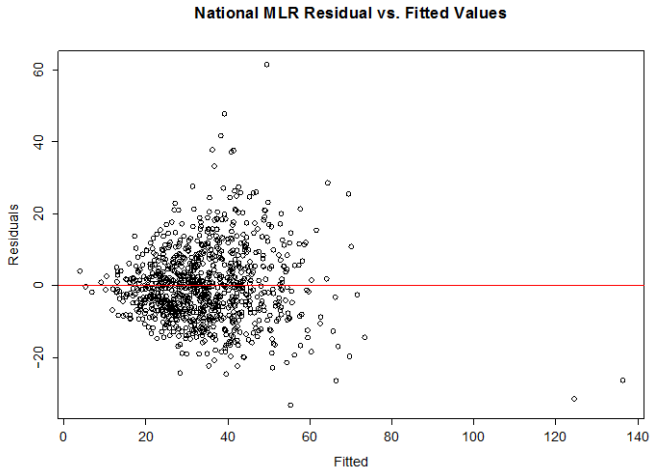


Figure 12. The residual versus fitted values plot validates the assumption that the error term ($\varepsilon$) has a constant variance as noted by the concentration of points. In this plot of the pre-transform fitted model, there exists a non-constant variance. Variance is minimal at lower fitted values and maximum at 55 accessions. Transformation of the response is one method of dealing with the non-constant variance (after Faraway, 2006). Figure 14 shows the results of the Box Cox test, which prescribe an ideal transformation (after Venables & Ripley, 2002).

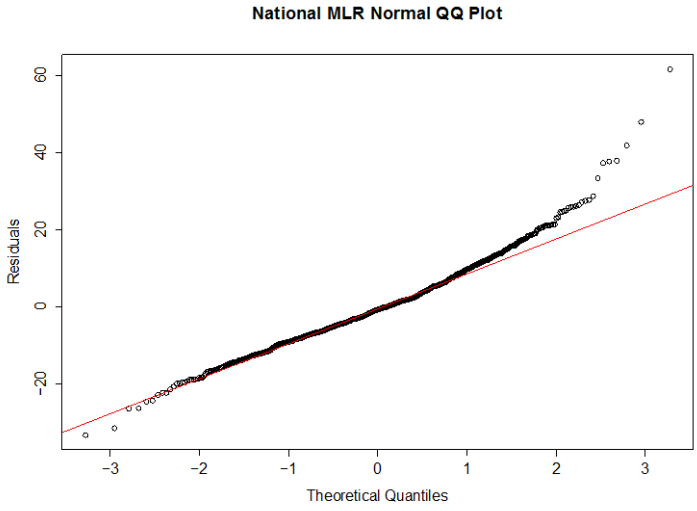


Figure 13. Normal QQ plot of the pre-transform fitted model indicates non-normal distribution of residuals; therefore, violating model assumptions.

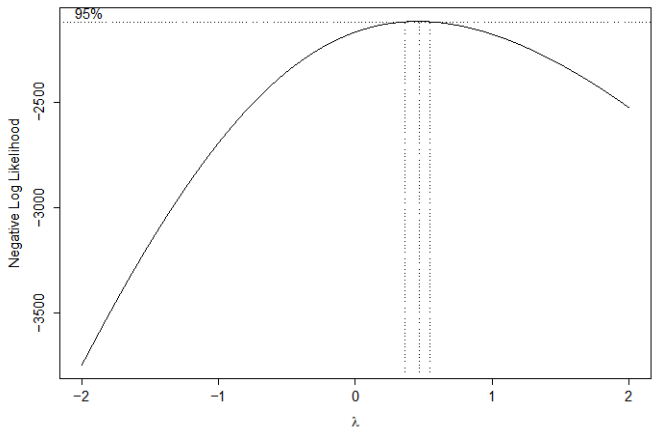


Figure 14. Interpretation of the Box Cox test reveals a transformation of square root. This is an ideal transformation when dealing with Poisson distributed count data (after R. Silvestrini, personal communication, 2014). Figure 14 supports this claim with $\lambda$ of .5, where lambda is the recommended power to which the response must be raised in order to achieve a normal response. Figure 15 shows the resultant variance following the transformation.

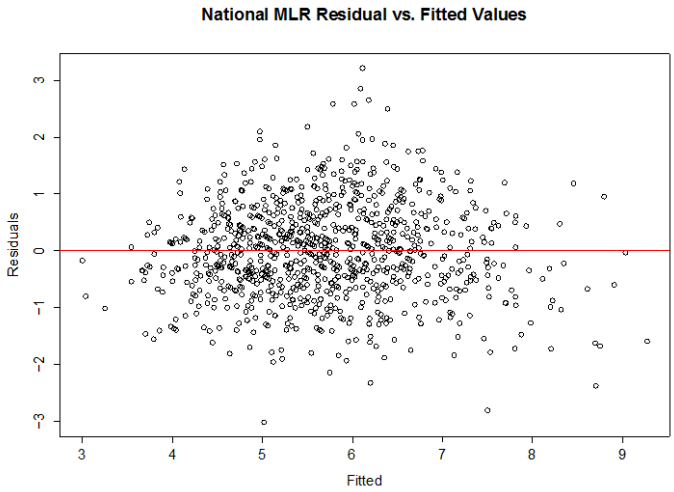


Figure 15. The post-transformation of the response to $\sqrt{y}$ indicates an improved fit meeting model assumptions.

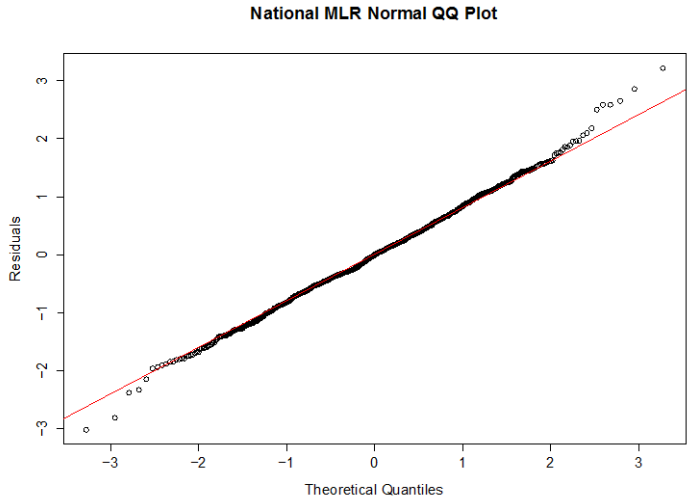


Figure 16. Improved normal QQ plot post-transformation. Shapiro-Wilk test confirmed normality with a p-value greater than .05 (after R Core Team, 2013).

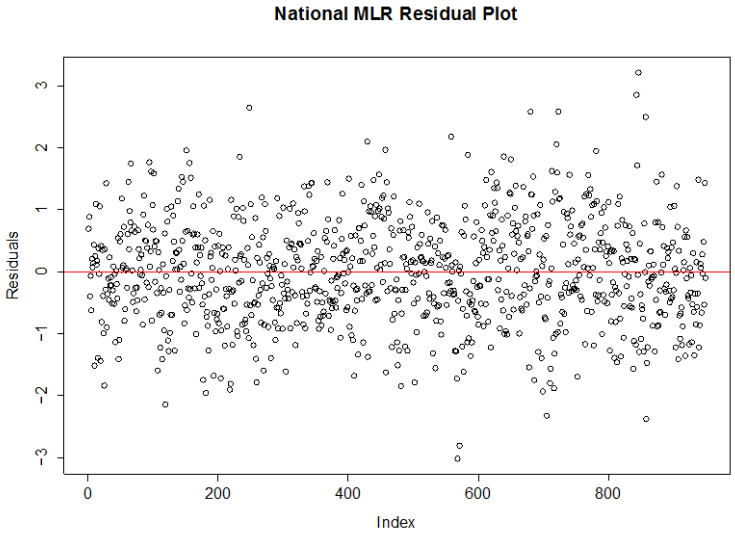


Figure 17. Residual plot interpretation of randomly distributed errors indicates that the errors are uncorrelated.

## B. EAST MODEL

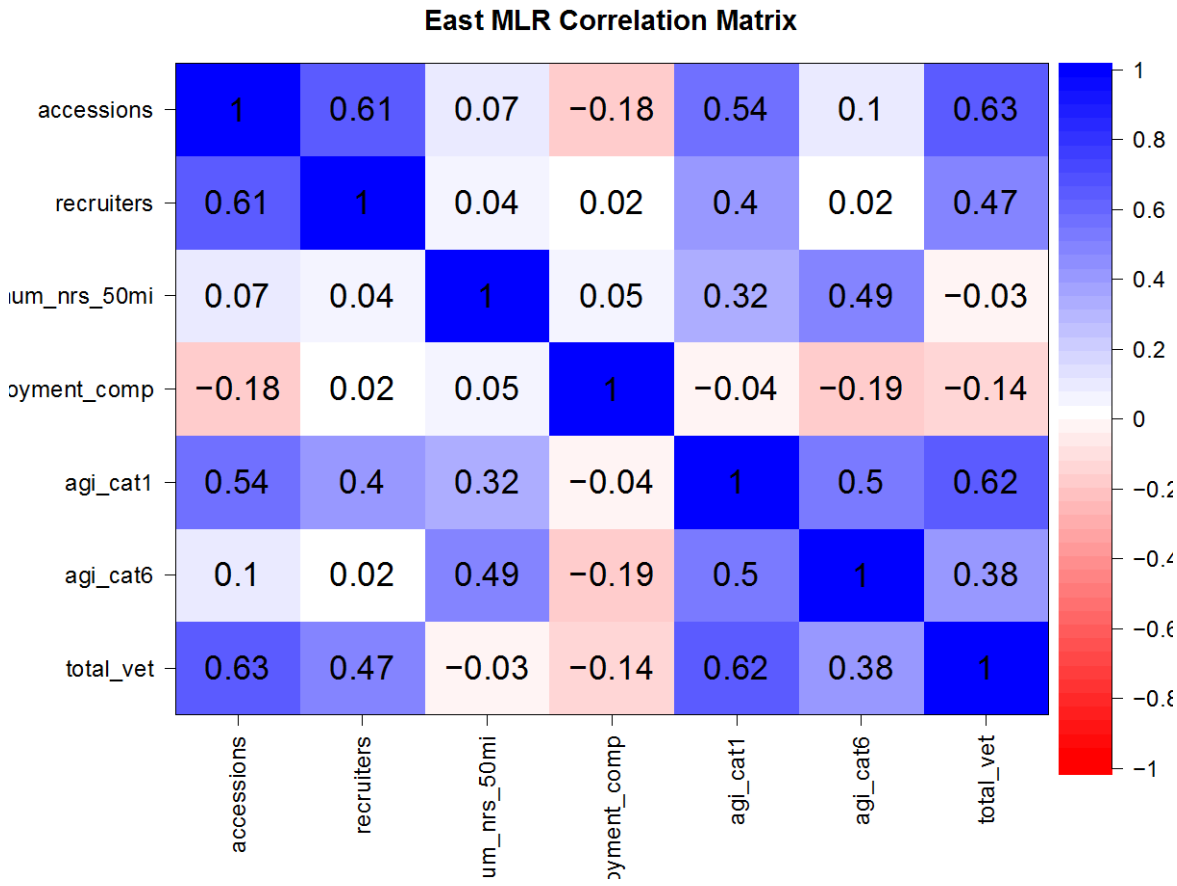


Figure 18. Correlation matrix showing the degree of correlation between the final subset of predictor variables and the response variable.

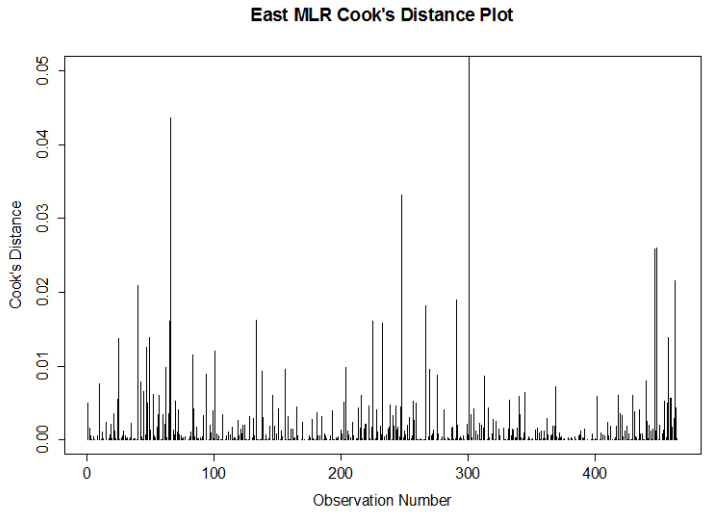


Figure 19. The fitted national model contained one outlier, as defined by Cook's distance > .05. Outlier was removed and model fit improved.

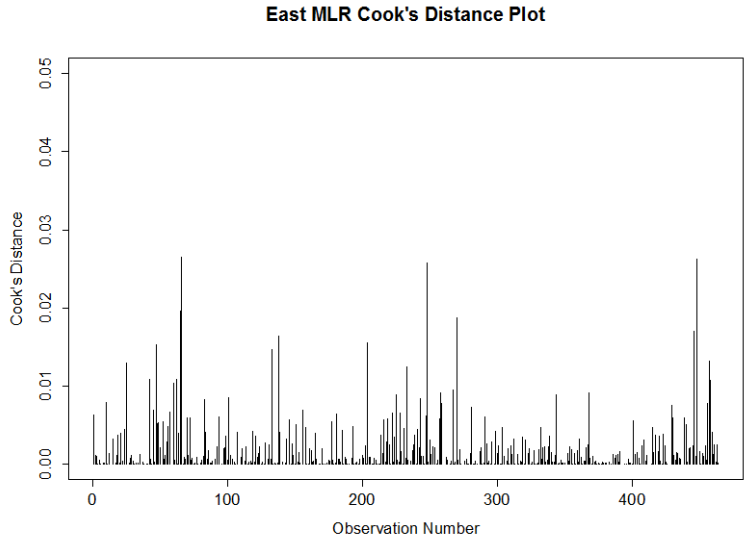


Figure 20. Plot showing the Cook's distance of all observations in final fitted model without outlier.

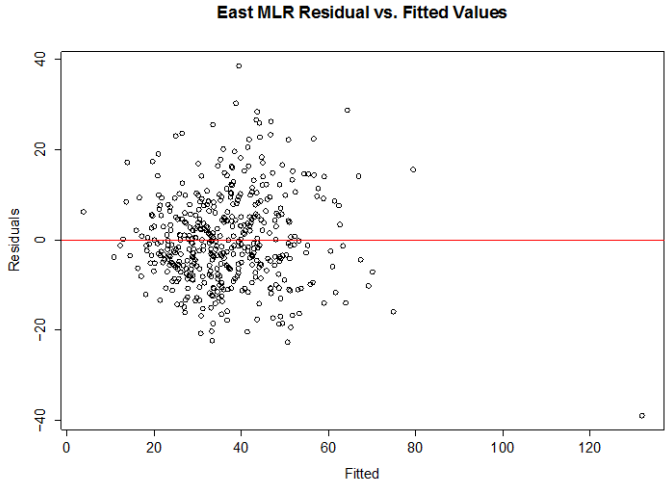


Figure 21. The residual versus fitted values plot validates the assumption that the error term ($\varepsilon$) has a constant variance. In this plot of the pre-transform fitted model, there exists a non-constant variance. Variance is minimal at lower fitted values and maximum at 40 accessions. Transformation of the response is one method of dealing with the non-constant variance (after Faraway, 2006). Figure 23 shows the results of the Box Cox test, which prescribe an ideal transformation (after Venables & Ripley, 2002).

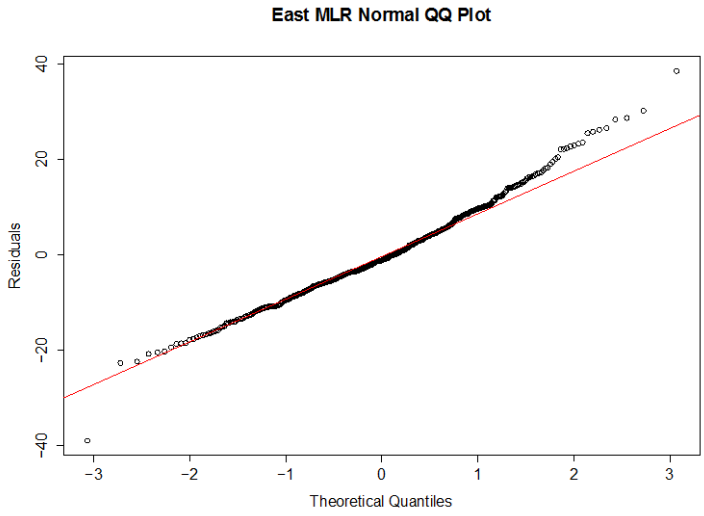


Figure 22. Normal QQ plot of the pre-transform fitted model indicates non-normal distribution of residuals; therefore, violating model assumptions.

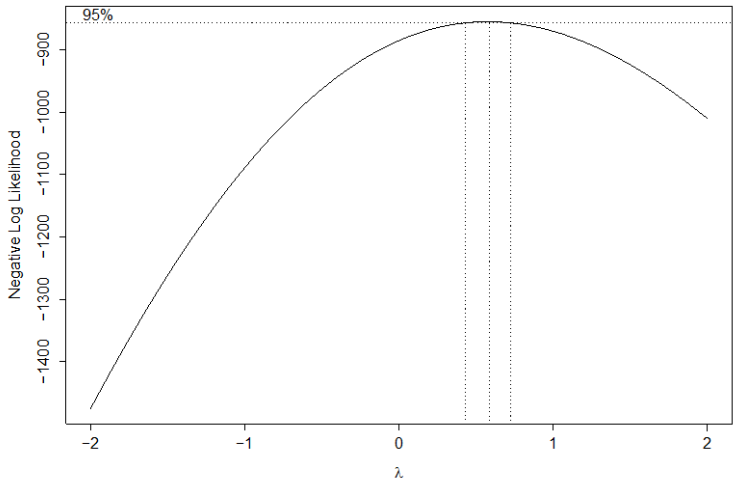


Figure 23. Interpretation of the Box Cox test reveals a transformation of square root. This is an ideal transformation when dealing with Poisson distributed count data (after R. Silvestrini, personal communication, 2014). Figure 23 supports this claim with $\lambda$ of .5, where lambda is the recommended power to which the response must be raised in order to achieve a normal response. Figure 24 shows the resultant variance following the transformation.

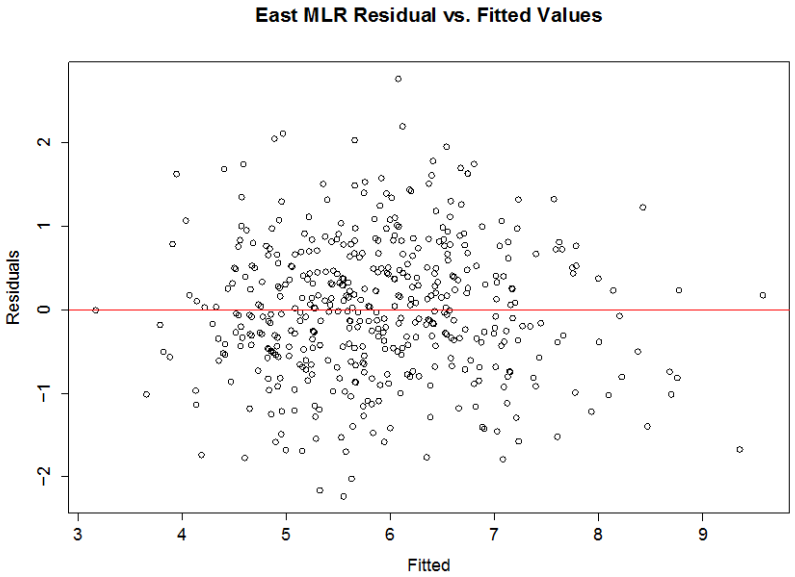


Figure 24. The post-transformation of the response to $\sqrt{y}$ indicates an improved fit meeting model assumptions.

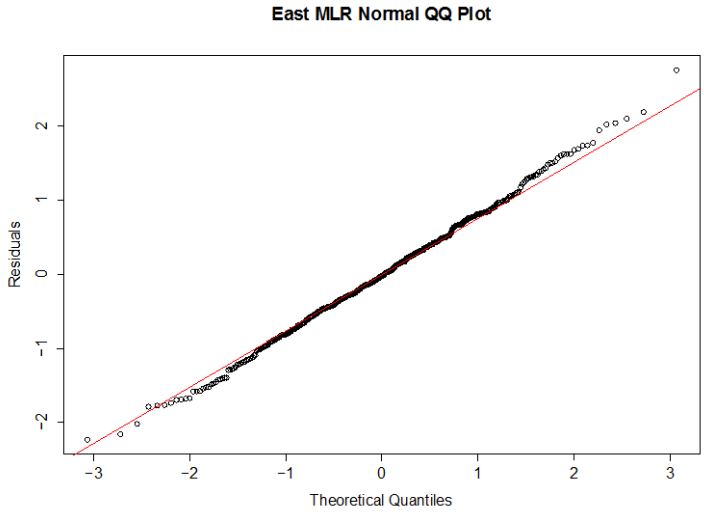


Figure 25. Improved normal QQ plot post-transformation. Shapiro-Wilk test confirmed normality with a p-value greater than .05 (after R Core Team, 2013).

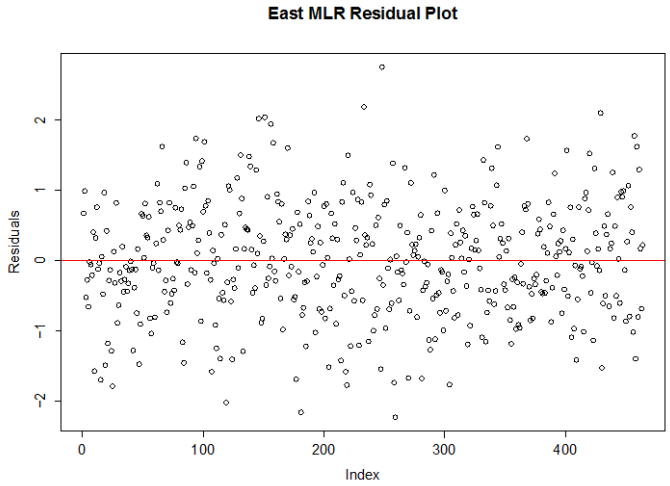


Figure 26. Residual plot interpretation of randomly distributed errors indicates that the errors are uncorrelated.

## C. WEST MODEL

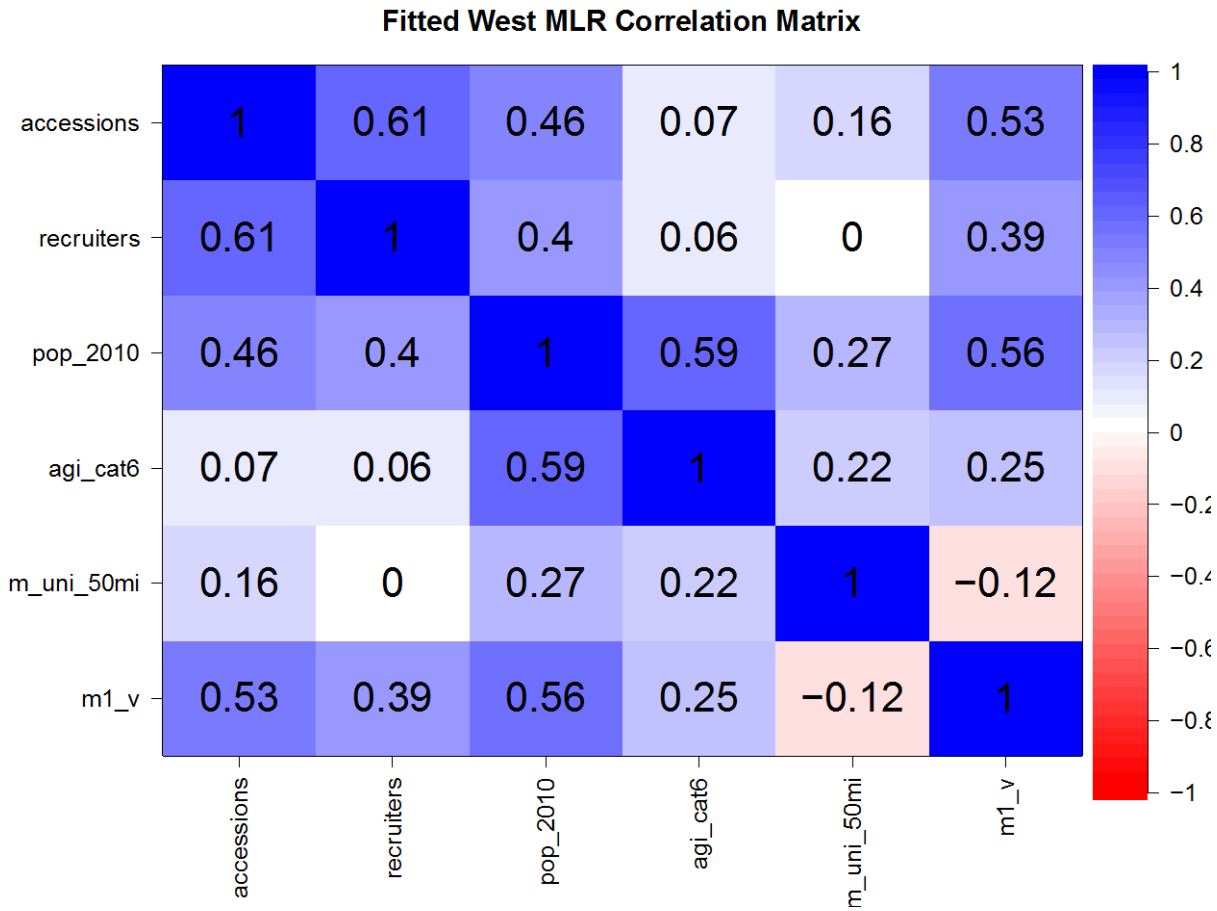


Figure 27. Correlation matrix showing the degree of correlation between the final subset of predictor variables and the response variable.

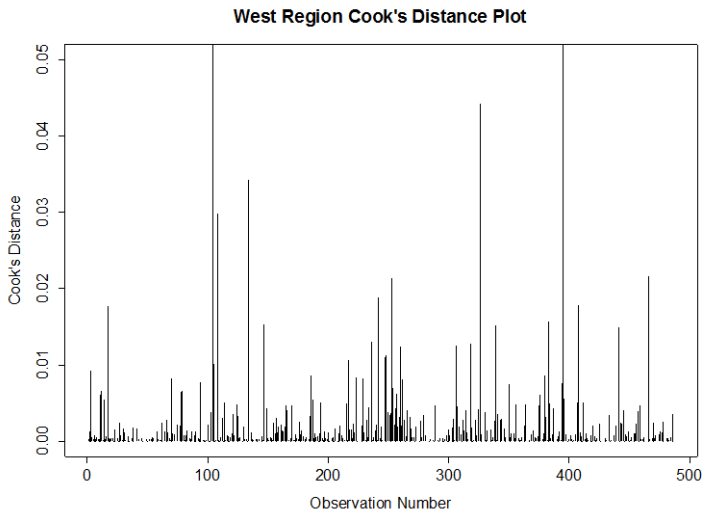


Figure 28. The fitted national model contained two outliers, as defined by Cook's distance > .05. Outliers were removed and model fit improved.

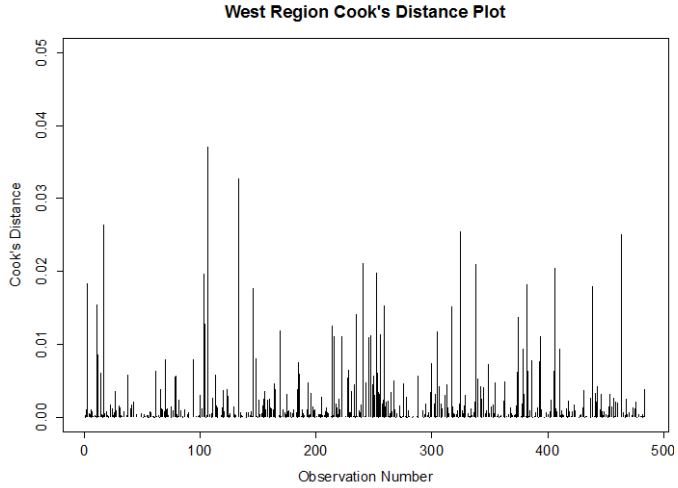


Figure 29. Plot showing the Cook's distance of all observations in final fitted model without outliers.

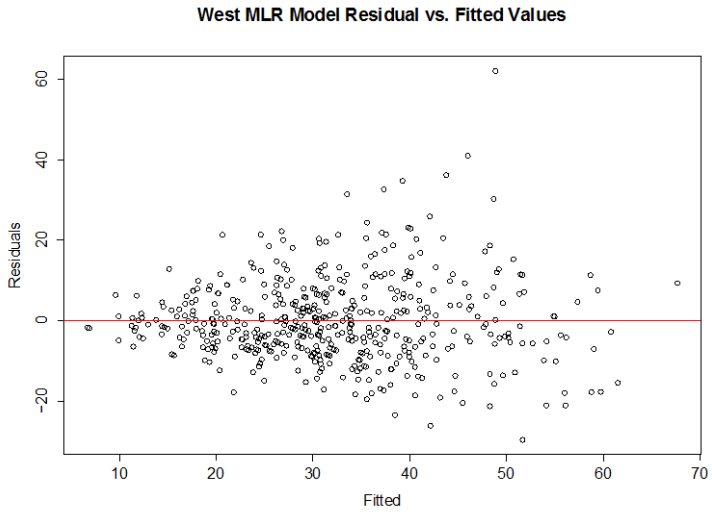


Figure 30. The residual versus fitted values plot validates the assumption that the error term ($\varepsilon$) has a constant variance. In this plot of the pre-transform fitted model, there exists a non-constant variance. Variance is minimal at lower fitted values and steadily increases with greater accessions to a maximum at 51 accessions. Transformation of the response is one method of dealing with the non-constant variance (after Faraway, 2006). Figure 32 shows the results of the Box Cox test, which prescribe an ideal transformation (after Venables & Ripley, 2002).

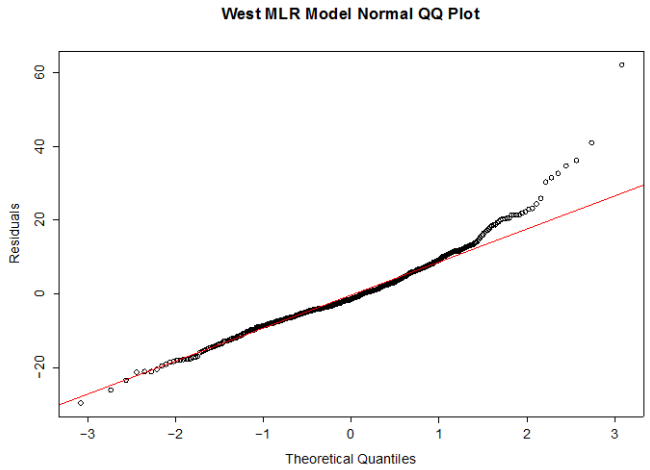


Figure 31. Normal QQ plot of the pre-transform fitted model indicates non-normal distribution of residuals; therefore, violating model assumptions.

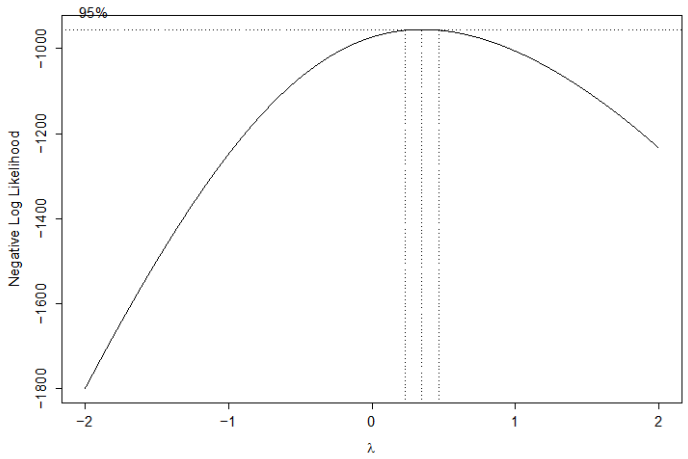


Figure 32. Interpretation of the Box Cox test reveals a transformation of cubed root. Figure 32 supports this claim with $\lambda$ of .34, where lambda is the recommended power to which the response must be raised in order to achieve a normal response. Figure 33 shows the resultant variance following the transformation.

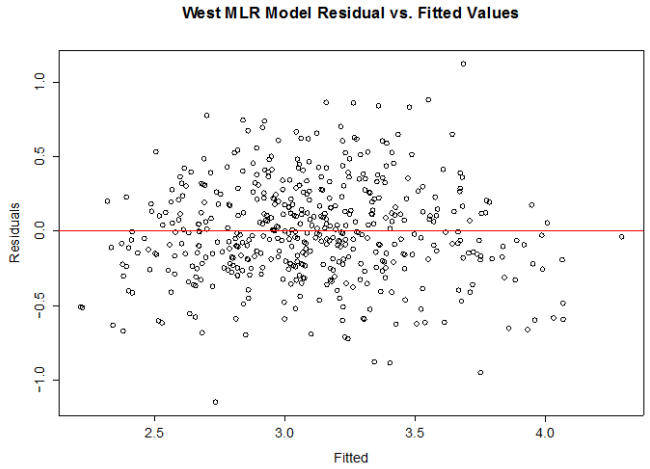


Figure 33. The post-transformation of the response to $\sqrt[3]{y}$ indicates an improved fit meeting model assumptions.

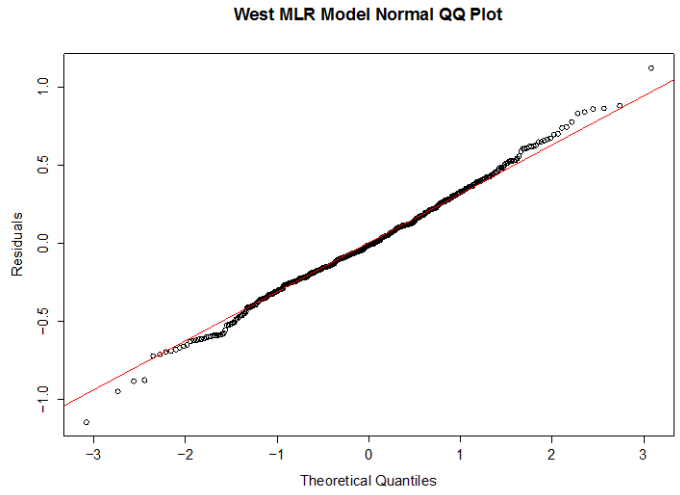


Figure 34. Improved normal QQ plot post-transformation. Shapiro-Wilk test confirmed normality with a p-value greater than .05 (aftre R Core Team, 2013).

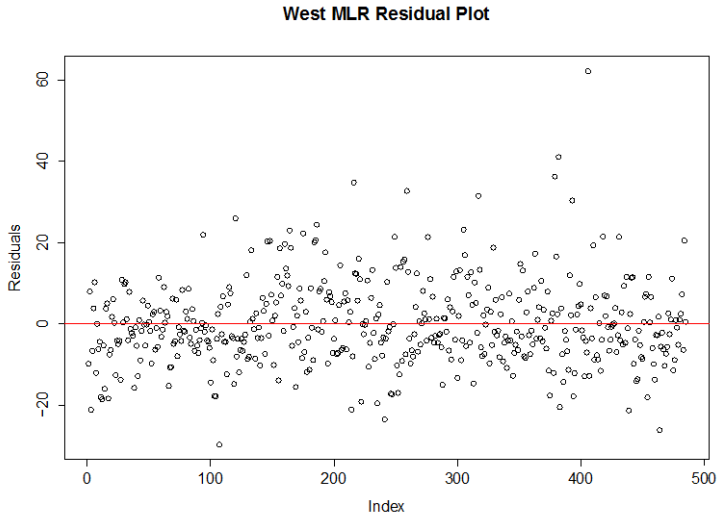


Figure 35. Residual plot interpretation of randomly distributed errors indicates that the errors are uncorrelated.

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX D: ZINB MODEL

Appendix D provides the reader with model variable summaries and east and west models.

Table 28. Summary of the model variables for the binomial component. "X" identifies the variable as present in the model.

| Binomial | National | East | West |
|---|---|---|---|
| average number of recruiters per year | X | X | |
| reports of violent crime | X | X | X |
| QMA | X | X | X |
| unemployment compensation | X | X | |
| AGI > $200,000 | X | X | X |
| number of Division I Universities within 50 miles of the ZIP code centroid | X | | |
| total number of universities in the ZIP code | X | X | |
| total number of veterans | X | X | |
| male veterans ages 18–34 | | | X |

Table 29. Summary of the model variables for the count component.

| Count | National | East | West |
|---|---|---|---|
| average number of recruiters per year | X | X | X |
| distance to NRS | X | X | X |
| number of NRSs within 50 miles | X | | X |
| reports of violent crime | X | X | X |
| population density | | | X |
| mean unemployment rate | X | | X |
| unemployment compensation | X | | X |
| AGI < $25,000 | X | X | X |
| AGI > $200,000 | X | X | X |
| distance to the nearest Division I university from the ZIP code centroid | | X | |
| number of Division I universities within 50 miles of the ZIP code centroid | X | | |
| total number of universities in the ZIP code | X | X | |
| total number of veterans | X | X | X |

Table 30. Summary view of the logistic component of the east ZINB model.

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 2.598 | 0.134 | 19.454 | 0.000 |
| average number of recruiters per year | –0.029 | 0.011 | –2.659 | 0.008 |
| reports of violent crime | –0.008 | 0.002 | –3.399 | 0.001 |
| QMA | –0.006 | 4e–04 | –13.649 | 0.000 |
| unemployment compensation | –0.033 | 0.009 | –3.558 | 0.000 |
| AGI > $200,000 | 0.002 | 3e–04 | 5.759 | 0.000 |
| total number of universities in the ZIP code | 1.086 | 0.283 | 3.839 | 0.000 |
| total number of veterans | –0.003 | 3e–04 | –8.626 | 0.000 |

Table 31. Summary view of the negative binomial component of the east ZINB model.

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | –0.354 | 0.027 | –12.917 | 0.000 |
| average number of recruiters per year | 0.012 | 0.001 | 10.400 | 0.000 |
| distance to NRS | –0.012 | 0.001 | –8.640 | 0.000 |
| reports of violent crime | 0.001 | 2e–04 | 4.254 | 0.000 |
| AGI < $25,000 | 7e–05 | 3e–06 | 27.638 | 0.000 |
| AGI > $200,000 | –1e–04 | 3e–05 | –5.679 | 0.000 |
| distance to the nearest Division I university from the ZIP code centroid | 0.002 | 0.001 | 3.446 | 0.001 |
| total number of universities in the ZIP code | –0.066 | 0.021 | –3.116 | 0.002 |
| total number of veterans | 4e–04 | 2e–05 | 21.467 | 0.000 |
| Log(theta) | 1.746 | 0.010 | 170.092 | 0.000 |

Table 32. Summary view of the logistic component of the west ZINB model.

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 2.102 | 0.065 | 32.117 | 0.000 |
| reports of violent crime | –0.032 | 0.004 | –7.052 | 0.000 |
| QMA | –0.004 | 5e–04 | –8.688 | 0.000 |
| AGI > $200,000 | 0.001 | 3e–04 | 3.551 | 0.000 |
| male veterans ages 18–34 | –0.043 | 0.005 | –9.117 | 0.000 |

Table 33. Summary view of the negative binomial component of the west ZINB model.

| Independent Variable | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | –0.365 | 0.050 | –7.369 | 0.000 |
| average number of recruiters per year | 0.014 | 0.001 | 14.019 | 0.000 |
| distance to NRS | –0.007 | 0.001 | –7.713 | 0.000 |
| number of NRSs within 50 miles | 0.011 | 0.001 | 8.690 | 0.000 |
| reports of violent crime | 0.004 | 0.001 | 4.744 | 0.000 |
| population density | –4e–05 | 1e–05 | –3.060 | 0.002 |
| mean unemployment rate | 0.017 | 0.005 | 3.132 | 0.002 |
| unemployment compensation | –0.010 | 0.003 | –2.968 | 0.003 |
| AGI < $25,000 | –5e–05 | 8e–06 | 6.081 | 0.000 |
| AGI > $200,000 | –8e–05 | 3e–05 | –3.113 | 0.002 |
| total number of veterans | 4e–04 | 2e–05 | 17.773 | 0.000 |
| Log(theta) | 1.478 | 0.015 | 100.558 | 0.000 |

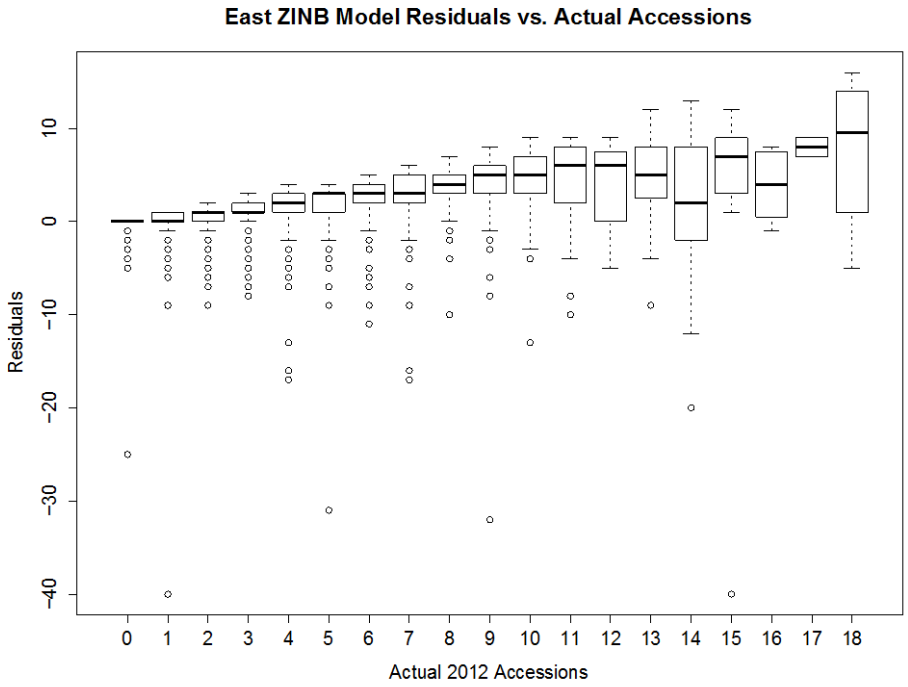


Figure 36. Plot of the residuals (actual accessions minus predicted accessions) versus the actual accessions for the east ZINB model, indicating a slight positive trend diverging from 0 residuals up to 12 accessions where the mean begins to re-converge with 0 residuals and eventually plateaus.

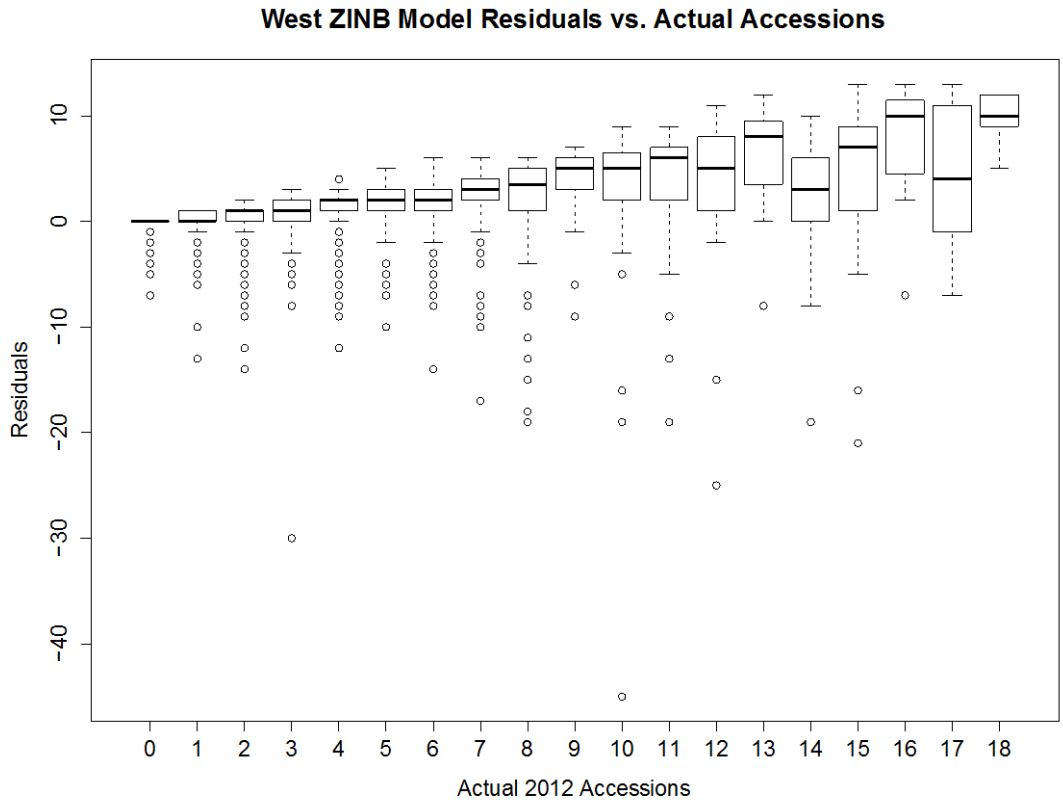


Figure 37. Plot of the residuals versus the actual accessions for the west ZINB model, indicating a slight positive trend diverging from 0 residuals. The bottom line is the ZINB model is very accurate at in ZIP codes with fewer than 5 accessions, which in the west could be very helpful for predicting zeros and accessions.

# APPENDIX E: R AND PYTHON SCRIPTS

Appendix E provides the reader with all R and PYTHON scripts used to clean and analyze data and fit MLR and ZINB models.

Table 34. ZIP code proportion PYTHON (after Jackson, 2015)

```
import math
from operator import itemgetter
import pandas
import scipy
import pdb
import networkx

#read in the data
zfile = pandas.read_csv('zip_data.csv')
ufile = pandas.read_csv('d1_data.csv')

#create the columns you want
zfile['nearest_uni']=''
zfile['dist_to_uni']=''
zfile['num_uni_50mi']=''

# Function will convert two pofloats to a distance
def calc_distance(lat1,lon1,lat2,lon2):
   rad_lat1 = lat1*2*math.pi/360
   rad_lon1 = lon1*2*math.pi/360
   rad_lat2 = lat2*2*math.pi/360
   rad_lon2 = lon2*2*math.pi/360
   #pdb.set_trace()
   dist_rad =
math.acos(math.sin(rad_lat1)*math.sin(rad_lat2)+math.cos(rad_lat1)*math.cos(rad_lat2)*math.cos(rad_l
on1–rad_lon2))
   dist_nm = ((180*60)/math.pi)*dist_rad
   return dist_nm

#create network
#digraph = networkx.Graph()

#fill the dist_to_uni column, num_uni_50mi
nearUni=[]
num_50mi=[]
for i,zipFrame in zfile.iterrows():
   tempNearUni = []
   for uni,uniFrame in ufile.iterrows():
      #pdb.set_trace()
      dist = calc_distance(zipFrame.zip_lat, zipFrame.zip_long, uniFrame.u_lat, uniFrame.u_long)
      tempNearUni.append(dist)
```

```
    #pdb.set_trace()
    value = scipy.amin(tempNearUni)
    nearUni.append(value)
    #pdb.set_trace()
    temp_num_50mi=[]
    for k in tempNearUni:
        #pdb.set_trace()
        if k <= 50:
            temp_num_50mi.append(k)
    number_o_unis = len(temp_num_50mi)
    #pdb.set_trace()
    num_50mi.append(number_o_unis)

dist_temp = scipy.array(nearUni)
num_unis_temp = scipy.array(num_50mi)

zfile['dist_to_uni'] = dist_temp
zfile['num_uni_50mi'] = num_unis_temp

#write to csv
zfile.to_csv('zip_to_uni_dists.csv')
```

Table 35. University data cleaning PYTHON and R scripts.

```
import math
from operator import itemgetter
import pandas
import scipy
import pdb
import networkx

#read in the data
zfile = pandas.read_csv('zip_data.csv')
ufile = pandas.read_csv('d1_data.csv')

#create the columns you want
zfile['nearest_uni']=''
zfile['dist_to_uni']=''
zfile['num_uni_50mi']=''

# Function will convert two pofloats to a distance
def calc_distance(lat1,lon1,lat2,lon2):
    rad_lat1 = lat1*2*math.pi/360
    rad_lon1 = lon1*2*math.pi/360
    rad_lat2 = lat2*2*math.pi/360
    rad_lon2 = lon2*2*math.pi/360
    #pdb.set_trace()
    dist_rad =
 math.acos(math.sin(rad_lat1)*math.sin(rad_lat2)+math.cos(rad_lat1)*math.cos(rad_lat2)*math.cos(rad
 _lon1–rad_lon2))
    dist_nm = ((180*60)/math.pi)*dist_rad
```

```
    return dist_nm

#create network
#digraph = networkx.Graph()

#fill the dist_to_uni column, num_uni_50mi
nearUni=[]
num_50mi=[]
for i,zipFrame in zfile.iterrows():
    tempNearUni = []
    for uni,uniFrame in ufile.iterrows():
        #pdb.set_trace()
        dist = calc_distance(zipFrame.zip_lat, zipFrame.zip_long, uniFrame.u_lat, uniFrame.u_long)
        tempNearUni.append(dist)
    #pdb.set_trace()
    value = scipy.amin(tempNearUni)
    nearUni.append(value)
    #pdb.set_trace()
    temp_num_50mi=[]
    for k in tempNearUni:
        #pdb.set_trace()
        if k <= 50:
            temp_num_50mi.append(k)
    number_o_unis = len(temp_num_50mi)
    #pdb.set_trace()
    num_50mi.append(number_o_unis)

dist_temp = scipy.array(nearUni)
num_unis_temp = scipy.array(num_50mi)

zfile['dist_to_uni'] = dist_temp
zfile['num_uni_50mi'] = num_unis_temp

#write to csv
zfile.to_csv('zip_to_uni_dists.csv')

### University Cleaning ###

ipeds <– read.csv("~/Desktop/NPS/Thesis/Data/University Data/ipeds_master.csv")
unique <– read.csv("~/Desktop/NPS/Thesis/Data/University Data/unique_u_zips.csv")
names(ipeds)
dim(ipeds)
names(unique)
dim(unique)


for (i in 1:nrow(unique)) {
  who <– which(unique$zip[i] == ipeds$zip)
  unique[i, "total_schools"] <- mean(ipeds[who,"total_schools"])
  unique[i, "enroll2013"] <- max(ipeds[who,"enroll2013"])
  unique[i, "enroll2013.L1000"] <- sum(ipeds[who,"enroll2013.L1000"])
```

```
  unique[i, "enroll2013.L5000"] <- sum(ipeds[who,"enroll2013.L5000"])
  unique[i, "enroll2013.L10000"] <- sum(ipeds[who,"enroll2013.L10000"])
  unique[i, "enroll2013.L20000"] <- sum(ipeds[who,"enroll2013.L20000"])
  unique[i, "enroll2013.G20000"] <- sum(ipeds[who,"enroll2013.G20000"])
  unique[i, "aid2013"] <- mean(ipeds[who,"aid2013"], na.rm=TRUE)
  unique[i, "in_state_2013"] <- mean(ipeds[who,"in_state_2013"], na.rm=TRUE)
  unique[i, "out_state_2013"] <- mean(ipeds[who,"out_state_2013"], na.rm=TRUE)

  unique[i, "enroll2012"] <- max(ipeds[who,"enroll2012"])
  unique[i, "enroll2012.L1000"] <- sum(ipeds[who,"enroll2012.L1000"])
  unique[i, "enroll2012.L5000"] <- sum(ipeds[who,"enroll2012.L5000"])
  unique[i, "enroll2012.L10000"] <- sum(ipeds[who,"enroll2012.L10000"])
  unique[i, "enroll2012.L20000"] <- sum(ipeds[who,"enroll2012.L20000"])
  unique[i, "enroll2012.G20000"] <- sum(ipeds[who,"enroll2012.G20000"])
  unique[i, "aid2012"] <- mean(ipeds[who,"aid2012"], na.rm=TRUE)
  unique[i, "in_state_2012"] <- mean(ipeds[who,"in_state_2012"], na.rm=TRUE)
  unique[i, "out_state_2012"] <- mean(ipeds[who,"out_state_2012"], na.rm=TRUE)

  unique[i, "enroll2011"] <- max(ipeds[who,"enroll2011"])
  unique[i, "enroll2011.L1000"] <- sum(ipeds[who,"enroll2011.L1000"])
  unique[i, "enroll2011.L5000"] <- sum(ipeds[who,"enroll2011.L5000"])
  unique[i, "enroll2011.L10000"] <- sum(ipeds[who,"enroll2011.L10000"])
  unique[i, "enroll2011.L20000"] <- sum(ipeds[who,"enroll2011.L20000"])
  unique[i, "enroll2011.G20000"] <- sum(ipeds[who,"enroll2011.G20000"])
  unique[i, "aid2011"] <- mean(ipeds[who,"aid2011"], na.rm=TRUE)
  unique[i, "in_state_2011"] <- mean(ipeds[who,"in_state_2011"], na.rm=TRUE)
  unique[i, "out_state_2011"] <- mean(ipeds[who,"out_state_2011"], na.rm=TRUE)
}
u_master <- merge (main, unique, by = "zip", all.x=TRUE, all.y=FALSE)

write.table(u_master, "~/Desktop/NPS/Thesis/Data/University Data/summed_u_master.csv", sep="\t")

ipeds[,("zip" == 32601)]
```

Table 36. ZIP to FIPS conversion and crime data cleaning scripts.

```
#Read in the file with the zip code populations

zips <- read.csv("~/Desktop/NPS/Thesis/Data/Crime Data/zip_pop.csv")
states <- read.csv("~/Desktop/NPS/Thesis/Data/Zip Conversion/states.csv")
names(zips)

state<-c("AK","CT","DC","HI","MA","RI")
zips2 <- subset(zips, state1==state)

zips1 <- subset(zips, state1=="AK")
for (i in 1:nrow(zips1)){
  zips1$sum[i]<-sum(zips1$zip_pop)
  zips1$prop_of_state[i]<-zips1$zip_pop[i]/zips1$sum[1]
  zips1$violent[i]<-states$violent[2]
```

```
}
zips2 <- subset(zips, state1=="CT")
for (i in 1:nrow(zips2)){
  zips2$sum[i]<-sum(zips2$zip_pop)
  zips2$prop_of_state[i]<-zips2$zip_pop[i]/zips2$sum[1]
}
zips3 <- subset(zips, state1=="DC")
for (i in 1:nrow(zips3)){
  zips3$sum[i]<-sum(zips3$zip_pop)
  zips3$prop_of_state[i]<-zips3$zip_pop[i]/zips3$sum[1]
}
zips4$sum <- 0
zips4$prop_of_state <- 0
zips4 <- subset(zips, state1=="HI")
for (i in 1:nrow(zips4)){
  zips4$sum[i]<-sum(zips4$zip_pop)
  zips4$prop_of_state[i]<-zips4$zip_pop[i]/zips4$sum[1]
}
sum(zips4$prop_of_state)
zips4$sum[1]
zips5 <- subset(zips, state1=="MA")
for (i in 1:nrow(zips5)){
  zips5$sum[i]<-sum(zips5$zip_pop)
  zips5$prop_of_state[i]<-zips5$zip_pop[i]/zips5$sum[1]
}
zips6 <- subset(zips, state1=="RI")
for (i in 1:nrow(zips6)){
  zips6$sum[i]<-sum(zips6$zip_pop)
  zips6$prop_of_state[i]<-zips6$zip_pop[i]/zips6$sum[1]
}

zips4state<-rbind(zips1,zips2,zips3,zips4,zips5,zips6)

names(zips4state)
zips4state$prop<-NULL
zips4state$fips<-NULL
zips4state$fips_pop<-NULL
zips4state$rsid<-NULL
zips4state$county<-NULL
zips4state$state2<-NULL

states<- read.csv("~/Desktop/NPS/Thesis/Data/Crime Data/state_2011.csv")

# Build a key for zips. It's not quite unique.
zips$Key <- paste (zips$state2, zips$county, sep=".")


#Build a key for the zip fips to group them into
zips$key <-paste(zips$fips, zips$zip, sep="_")
```

```
#Loop to sum all zip code populations for each fips and prop each zip makes zip code has in that county
zips$state_pop <- 0
zips$prop_of_state <- 0

for (i in 1:nrow(states)) {
   states$state_pop <- sum(zips$zip_pop)tapply(zips$state1, zips$zip_pop, FUN = (zip_pop[i]/sum),
 simplify = TRUE)
   }
?sum
tapply(1:3, zips$fip[1:3], sum)

state_pop = by(zips,zips$state2,function(x) {sum(x$zip_pop)})
state_pop
statepop<-as.matrix(state_pop)
typeof(state_pop)
popFip[2]
zips$sum <- 0
zips$prop <- 0
for (i in 1:nrows(zips)) {
  popFip[zips$fips[i]
  zips$prop[i] <- zips$pop[i]/popFip[zips$fips[i]]
}

popFip = by(zips,zips$fips,function(x) {sum(x$pop)})
popFip[2]
zips$sum <- 0
zips$prop <- 0


sums<-as.matrix(popFip)
dim(sums)
for (i in 1:nrows(zips)) {
  popFip[zips$fips[i]
  zips$sum[i] <- popFip[zips$fips[i]]
}


write.table(zips, "~/Desktop/NPS/Thesis/Data/Crime Data/zips_prop.csv", sep="\t")

zips[zips$fips =="25027",]
sum(zips$prop>0)
tail(zips)
head(zips)

max(zips$fip)

write.table(statepop, "~/Desktop/NPS/Thesis/Data/Crime Data/2010_state_pop.csv", sep="\t")

write.table(sums, "~/Desktop/NPS/Thesis/Data/Crime Data/countysums.csv", sep="\t")
names(states)
zips4state$wtd_violent<-0
```

```
zips4state$wtd_non_violent<-0
zips4state$violent<-0
zips4state$non_violent<-0
for (i in 1:nrows(zips4state)) {
  for zips4state$state1 == states$state2 {


  }
      zips$sum[i] <- popFip[zips$fips[i]]
}

names(zips4state)
write.table(zips4state, "~/Desktop/NPS/Thesis/Data/Crime Data/zips4state.csv", sep="\t")
zips4state$zips4state<-NULL
zips4state$wtd_violent<-NULL
zips4state$non_violent<-NULL
zips4state$wtd_non_violent<-NULL



# Distribution of crime incidents by ZIP.
#
# 1.) Read in crime data. I highlighted A1:D2578 from
# 2011_crime.csv and copied it to the clipboard. Then:
#
crime <- read.csv("~/Desktop/NPS/Thesis/Data/Crime Data/2011_crime.csv")

# Lose trailing spaces in the state and county columns.
crime$state <- sub (" +$", "", crime$state)
crime$county <- sub (" +$", "", crime$county)
# Change both kinds of crime to be numeric, first removing commas
crime$violent     <- as.numeric (sub (",", "", crime$violent))
crime$non_violent <- as.numeric (sub (",", "", crime$non_violent))

# Build a key for crime. It's not quite unique.
crime$Key <- paste (crime$state, crime$county, sep=".")

c2 <- data.frame (Key = sort (unique (crime$Key)),
   violent = tapply (crime$violent, crime$Key, sum),
   non_violent = tapply (crime$non_violent, crime$Key, sum), stringsAsFactors=FALSE)

# Check that this worked by matching row names to Keys.
all (c2$Key == row.names (c2))

#
# 2.) Read in the proportion.master
#
prop <- read.csv("~/Desktop/NPS/Thesis/Data/Crime Data/proportion_master.csv")
#
# 3.) Build a key for both files.
#
prop$Key  <- paste (prop$state,  prop$master_county, sep=".")
```

```
#
# 4.) Prop changes. First remove leading spaces from master_state.
#
prop$master_state <- sub ("^ +", "", prop$master_state)
# Handle Baltimore, St. Louis, and five kinds of Virginia

prop[prop$state == "MARYLAND"
  & prop$master_county == "Baltimore City", "master_county"] <- "Baltimore"
prop[prop$state == "MARYLAND"
  & prop$master_county == "Baltimore County", "master_county"] <- "Baltimore"
prop[prop$state == "MISSOURI"
  & prop$master_county == "St. Louis City", "master_county"] <- "St. Louis"
prop[prop$state == "MISSOURI"
  & prop$master_county == "St. Louis County", "master_county"] <- "St. Louis"
# Bedford is a little different
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Bedford County", "master_county"] <- "Bedford"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Fairfax County", "master_county"] <- "Fairfax"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Fairfax City", "master_county"] <- "Fairfax"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Franklin County", "master_county"] <- "Franklin"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Franklin City", "master_county"] <- "Franklin"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Richmond County", "master_county"] <- "Richmond"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Richmond City", "master_county"] <- "Richmond"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Roanoke County", "master_county"] <- "Roanoke"
prop[prop$state == "VIRGINIA"
  & prop$master_county == "Roanoke City", "master_county"] <- "Roanoke"
#
# Change "Dona Ana" with a tilde to "Dona Ana" without
#
prop$master_county[grep (" Ana", prop$master_county)] <- "Dona Ana"
#
# Rebuild the key
#
prop$Key  <- paste (prop$state,  prop$master_county, sep=".")
#
# Read in the set of changes to make to "crime"
#
#changers <- read.table (“clipboard”, sep=”\t”, header=T)
#changers <- read.table(file = "C:/Desktop/NPS/Thesis/Data/Crime Data/Changers.txt", sep="\t",
 header=T)
changers
#crimehelp(strsplit)
#, sep="\t", header=T))
changers <- changers[,c("Original.Crime", "New", "Operation")] # drop empties
```

```
changers <- changers[-grep ("^PROP", changers$Operation),]

keys.to.change <- c2$Key[match (changers$Original.Crime, c2$Key)]
new.values <- changers[match (keys.to.change, changers$Original.Crime), "New"]
c2$Key[match (keys.to.change, c2$Key)] <- new.values
#
# Delete "MARYLAND.Baltimore County" (no crimes reported; we assigned them to
# Baltimore City") and "TEXAS.King" (pop. 286, one nonviolent crime).
#
c2 <- c2[!is.element (c2$Key, c("MARYLAND.Baltimore County", "TEXAS.King")),]

#
# Now see who matches!
#
x<-table (is.element (c2$Key, prop$Key))
x
#table(x)["FALSE"]
c2[!is.element(c2$Key,prop$Key),"Key"]

#
prop$WtdViolent <- 0; prop$WtdNon <- 0
#
# Loop!
#
for (i in 1:nrow(c2)) {
   who <- which (prop$Key == c2$Key[i])
   prop[who, "WtdViolent"] <- c2[i,"violent"]     * prop[who, "prop_of_county"]
   prop[who, "WtdNon"]     <- c2[i,"non_violent"] * prop[who, "prop_of_county"]
}
```

Table 37. IRS data cleaning scripts.

```
### IRS Data Cleaning ###


#Import all years' data

irs2012 <- read.csv("~/Desktop/NPS/Thesis/Data/Economic/Salary/2012zipcode/12zpallagi.csv")
irs2012<-irs2012[7:102,]
names(irs2012)
irs2012_m <- read.csv("~/Desktop/NPS/Thesis/Data/Economic/Salary/2012zipcode/irs_master_zip.csv")
irs2012_m<-irs2012_m[11128:11146,]

irs2012 <- read.csv("~/Desktop/NPS/Thesis/Data/Economic/Salary/2012zipcode/12zpallagi.csv")
irs2012b <- read.csv("~/Desktop/NPS/Thesis/Data/Economic/Salary/2012zipcode/12zpallnoagi.csv")

names(irs2012)
dim(irs2012)
names(irs2012_m)
dim(irs2012_m)
```

```
## Replace all the 9999 and 0.0001 with NA to not mess up the calculations
#within(irs2012, levels(agi)[levels(agi) == 0.0001]<- "NA")
#within(irs2012, levels(a_pensions)[levels(a_pensions) == 0.0001]<- "NA")

irs2012[irs2012==0.0001]<- NA
irs2012[irs2012==99999]<- NA


names(irs2012)
names(irs2012_m)

#
# IRS action
#
# 1.) irs2012 is the origina data,
irs2012_m <- irs2012
for (i in 1:nrow(irs2012)) {
  nm <- paste("AGI_STUB", i, sep="")
  irs2012_m[nm] <- irs2012[irs2012$AGI_STUB == i,"N1"]
}

?ifelse
include <- irs2012_m$zip %in% irs2012$ZIPCODE
include
dim(exclude)
nm
for (i in 1:nrow(irs2012_m)){
  test<-t(irs2012[who,"N1"])

irs2012_m$six_fig <- 0

for (i in 1:nrow(irs2012_m)) {
  who <- which(irs2012_m$zip[i] == irs2012$zipcode)
  irs2012_m[i,  "avg_agi"]  <-  sum(irs2012[who,"A00100"],  na.rm=TRUE)*(1000)/sum(irs2012[who,"N1"],
na.rm=TRUE)
  irs2012_m[i,                "unemployment_comp"]                <-                (sum(irs2012[who,"N02300"],
na.rm=TRUE))/(sum(irs2012[who,"N1"], na.rm=TRUE))
  irs2012_m[i,                 "per_returns_taxable"]                <-                (sum(irs2012[who,"N04800"],
na.rm=TRUE))/(sum(irs2012[who,"N1"], na.rm=TRUE))
  irs2012_m[i,                "pensions_annuities_agi"]                <-                (sum(irs2012[who,"A01700"],
na.rm=TRUE))/(sum(irs2012[who,"A04800"], na.rm=TRUE))
  irs2012_m[i, "agi_cat1"] <- (irs2012[who,"N1"])[1]
  irs2012_m[i, "agi_cat2"] <- (irs2012[who,"N1"])[2]
  irs2012_m[i, "agi_cat3"] <- (irs2012[who,"N1"])[3]
  irs2012_m[i, "agi_cat4"] <- (irs2012[who,"N1"])[4]
  irs2012_m[i, "agi_cat5"] <- (irs2012[who,"N1"])[5]
  irs2012_m[i, "agi_cat6"] <- (irs2012[who,"N1"])[6]
  irs2012_m[i, "six_fig"] <- sum(irs2012[who,"N1"][5],irs2012[who,"N1"][6],na.rm=TRUE)
}
```

```
write.table(irs2012_m,
"~/Desktop/NPS/Thesis/Data/Economic/Salary/2012zipcode/irs2012_aggregated_master.csv", sep="\t")


x=c(1,2,3,4)
typeof(x)
sum(x[1],x[2])
irs2012_m[i, "agi_cat1"] <- (irs2012[who,"N1"])[1]
t <- (irs2012[who,"N1"])[1]
typeof(t)
t[1]
t[2]
t(t)

  irs2012_m[i, "enroll2013.L5000"] <- sum(irs2012[who,"enroll2013.L5000"])
  irs2012_m[i, "enroll2013.L10000"] <- sum(irs2012[who,"enroll2013.L10000"])
  irs2012_m[i, "enroll2013.L20000"] <- sum(irs2012[who,"enroll2013.L20000"])
  irs2012_m[i, "enroll2013.G20000"] <- sum(irs2012[who,"enroll2013.G20000"])
  irs2012_m[i, "aid2013"] <- mean(irs2012[who,"aid2013"], na.rm=TRUE)
  irs2012_m[i, "in_state_2013"] <- mean(irs2012[who,"in_state_2013"], na.rm=TRUE)
  irs2012_m[i, "out_state_2013"] <- mean(irs2012[who,"out_state_2013"], na.rm=TRUE)

  irs2012_m[i, "enroll2012"] <- max(irs2012[who,"enroll2012"])
  irs2012_m[i, "enroll2012.L1000"] <- sum(irs2012[who,"enroll2012.L1000"])
  irs2012_m[i, "enroll2012.L5000"] <- sum(irs2012[who,"enroll2012.L5000"])
  irs2012_m[i, "enroll2012.L10000"] <- sum(irs2012[who,"enroll2012.L10000"])
  irs2012_m[i, "enroll2012.L20000"] <- sum(irs2012[who,"enroll2012.L20000"])
  irs2012_m[i, "enroll2012.G20000"] <- sum(irs2012[who,"enroll2012.G20000"])
  irs2012_m[i, "aid2012"] <- mean(irs2012[who,"aid2012"], na.rm=TRUE)
  irs2012_m[i, "in_state_2012"] <- mean(irs2012[who,"in_state_2012"], na.rm=TRUE)
  irs2012_m[i, "out_state_2012"] <- mean(irs2012[who,"out_state_2012"], na.rm=TRUE)

  irs2012_m[i, "enroll2012"] <- max(irs2012[who,"enroll2012"])
  irs2012_m[i, "enroll2012.L1000"] <- sum(irs2012[who,"enroll2012.L1000"])
  irs2012_m[i, "enroll2012.L5000"] <- sum(irs2012[who,"enroll2012.L5000"])
  irs2012_m[i, "enroll2012.L10000"] <- sum(irs2012[who,"enroll2012.L10000"])
  irs2012_m[i, "enroll2012.L20000"] <- sum(irs2012[who,"enroll2012.L20000"])
  irs2012_m[i, "enroll2012.G20000"] <- sum(irs2012[who,"enroll2012.G20000"])
  irs2012_m[i, "aid2012"] <- mean(irs2012[who,"aid2012"], na.rm=TRUE)
  irs2012_m[i, "in_state_2012"] <- mean(irs2012[who,"in_state_2012"], na.rm=TRUE)
  irs2012_m[i, "out_state_2012"] <- mean(irs2012[who,"out_state_2012"], na.rm=TRUE)
}
u_master <- merge (main, irs2012_m, by = "zip", all.x=TRUE, all.y=FALSE)

write.table(u_master, "~/Desktop/NPS/Thesis/Data/University Data/summed_u_master.csv", sep="\t")
```

Table 38. Veteran data cleaning scripts.

```
#Read in the raw veteran data and the zips by fip proportion data

vets <- read.csv("~/Desktop/NPS/Thesis/Data/Veteran Data/vet_data.csv")
prop <- read.csv("~/Desktop/NPS/Thesis/Data/Veteran Data/prop_data_vet.csv")


#create new columns for the new wtd data
#Legend for the age categories
#age_cat  male      female        male_veteran        female_veteran     male_nonvet         female_nonvet
#18-34          m1     f1        m1_v            f1_v            m1_nv      f1_nv
#35-54          m2     f2        m2_v            f2_v            m2_nv      f2_nv
#55-64          m3     f3        m3_v            f3_v            m3_nv      f3_nv
#65-74          m4     f4        m4_v            f4_v             m4_nv            f4_nv
#75+      m5   f5        m5_v            f5_v            m5_nv      f5_nv


prop$tot <- 0; prop$totv <- 0; prop$totnv <- 0

prop$totm <- 0; prop$totmv <- 0; prop$totmnv <- 0
prop$m1 <- 0; prop$m1v <- 0; prop$m1nv <- 0
prop$m2 <- 0; prop$m2v <- 0; prop$m2nv <- 0
prop$m3 <- 0; prop$m3v <- 0; prop$m3nv <- 0
prop$m4 <- 0; prop$m4v <- 0; prop$m4nv <- 0
prop$m5 <- 0; prop$m5v <- 0; prop$m5nv <- 0

prop$totf <- 0; prop$totfv <- 0; prop$totfnv <- 0
prop$f1 <- 0; prop$f1v <- 0; prop$f1nv <- 0
prop$f2 <- 0; prop$f2v <- 0; prop$f2nv <- 0
prop$f3 <- 0; prop$f3v <- 0; prop$f3nv <- 0
prop$f4 <- 0; prop$f4v <- 0; prop$f4nv <- 0
prop$f5 <- 0; prop$f5v <- 0; prop$f5nv <- 0
#
# Loop!
names_list <- names(prop)
for name in names_list[7:43] {
 for (i in 1:nrow(prop))
   who <- which(vets$fips == prop$fips[i])
   if (length (who) > 0)
    prop[i, "tot"] <- vets[who,"total_pop"] * prop[i, "prop"]
#     prop[i, "totv"] <- vets[who,"total_vet"] * prop[i, "prop"]
#     prop[i, "totnv"] <- vets[who,"total_nonvet"] * prop[i, "prop"]
#
#     prop[i, "totm"] <- vets[who,"tot_m"] * prop[i, "prop"]
#     prop[i, "totmv"] <- vets[who,"tot_m_vet"] * prop[i, "prop"]
#     prop[i, "totmnv"] <- vets[who,"tot_m_nonvet"] * prop[i, "prop"]
#
#     prop[i, "m1"] <- vets[who,"m1"] * prop[i, "prop"]
#     prop[i, "m1v"] <- vets[who,"m1_v"] * prop[i, "prop"]
#     prop[i, "m1nv"] <- vets[who,"m1_nv"] * prop[i, "prop"]
#
```

```
#     prop[i, "m2"] <- vets[who,"m2"] * prop[i, "prop"]
#     prop[i, "m2v"] <- vets[who,"m2_v"] * prop[i, "prop"]
#     prop[i, "m2nv"] <- vets[who,"m2_nv"] * prop[i, "prop"]
#
#     prop[i, "m3"] <- vets[who,"m3"] * prop[i, "prop"]
#     prop[i, "m3v"] <- vets[who,"m3_v"] * prop[i, "prop"]
#     prop[i, "m3nv"] <- vets[who,"m3_nv"] * prop[i, "prop"]
#
#     prop[i, "m4"] <- vets[who,"m4"] * prop[i, "prop"]
#     prop[i, "m4v"] <- vets[who,"m4_v"] * prop[i, "prop"]
#     prop[i, "m4nv"] <- vets[who,"m4_nv"] * prop[i, "prop"]
#
#     prop[i, "m5"] <- vets[who,"m5"] * prop[i, "prop"]
#     prop[i, "m5v"] <- vets[who,"m5_v"] * prop[i, "prop"]
#     prop[i, "m5nv"] <- vets[who,"m5_nv"] * prop[i, "prop"]
#     prop[i, "totf"] <- vets[who,"tot_f"] * prop[i, "prop"]
#     prop[i, "totfv"] <- vets[who,"tot_f_vet"] * prop[i, "prop"]
#     prop[i, "totfnv"] <- vets[who,"tot_f_nonvet"] * prop[i, "prop"]
#
#     prop[i, "f1"] <- vets[who,"f1"] * prop[i, "prop"]
#     prop[i, "f1v"] <- vets[who,"f1_v"] * prop[i, "prop"]
#     prop[i, "f1nv"] <- vets[who,"f1_nv"] * prop[i, "prop"]
#
#     prop[i, "f2"] <- vets[who,"f2"] * prop[i, "prop"]
#     prop[i, "f2v"] <- vets[who,"f2_v"] * prop[i, "prop"]
#     prop[i, "f2nv"] <- vets[who,"f2_nv"] * prop[i, "prop"]
#
#     prop[i, "f3"] <- vets[who,"f3"] * prop[i, "prop"]
#     prop[i, "f3v"] <- vets[who,"f3_v"] * prop[i, "prop"]
#     prop[i, "f3nv"] <- vets[who,"f3_nv"] * prop[i, "prop"]
#
#     prop[i, "f4"] <- vets[who,"f4"] * prop[i, "prop"]
#     prop[i, "f4v"] <- vets[who,"f4_v"] * prop[i, "prop"]
#     prop[i, "f4nv"] <- vets[who,"f4_nv"] * prop[i, "prop"]
#
#     prop[i, "f5"] <- vets[who,"f5"] * prop[i, "prop"]
#     prop[i, "f5v"] <- vets[who,"f5_v"] * prop[i, "prop"]
#     prop[i, "f5nv"] <- vets[who,"f5_nv"] * prop[i, "prop"]
}



#
# Do that thing with vets and prop. Assumes fips is sorted in both.
#
overlappers <- vets$fips[vets$fips %in% prop$fips] # still in order
overlappers
fip.lens <- table (prop$fips)[as.character (overlappers)]
v2 <- vets[vets$fips %in% overlappers,]

sam <- v2[rep (1:nrow(v2), fip.lens),-(1:2)] * prop$prop[prop$fips %in% overlappers]
```

```
sam <- data.frame (fips = v2[rep (1:nrow(v2), fip.lens), "fips"], zip = prop$zip[prop$fips %in% overlappers],
 sam)
row.names (sam) <- 1:nrow(sam)

p2 <- prop[,1:4]
p2 <- merge (p2, sam, by = "zip", all.x=TRUE, all.y=FALSE)
p2$vet_fips <- NULL
p2[is.na (p2)] <- 0


write.table(p2, "~/Desktop/NPS/Thesis/Data/Veteran Data/vets_wtd.csv", sep="\t")

x<-table (is.element (p2$zip, prop$zip))
x
33179-33083
#table(x)["FALSE"]
prop[!is.element(prop$zip,p2$zip),"zip"]
p2[!is.element(p2$zip,prop$zip),"zip"]

which(duplicated(prop$zip))
d <- p2$zip
which(d == 2861
```

Table 39. Scripts for final cleaning and conversion from ZIP to station level of all data sets.

```
### Zip to Station ###

###  Aggregate 2011 data
final.data11 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2011.csv")
###  Aggregate 2012 data
final.data12 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2012.csv")
###  Aggregate 2012 data
final.data13 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2013.csv")

?by
rsid.accessions <- by(final.data11$accessions, final.data11$rsid, sum)

typeof(rsid.accessions)
sort.list(rsid.accessions, decreasing =FALSE)

# Another way

# length(final.data11$accessions)
# length(final.data11$rsid)
#

#
# length(tapply(final.data11$accessions, final.data11$rsid, sum))
y<-tapply(final.data11$accessions, final.data11$rsid, sum)
y
# typeof(y)
```

```
dist <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/zip_to_nrs.csv")
# Matrix of column names and operations
rsid.dist <- tapply(dist$dist_to_nrs, dist$rsid, median)
d<-as.matrix(rsid.dist)
rsid.num <- tapply(dist$num_nrs_50mi, dist$rsid, median)
n<-as.matrix(rsid.num)
t<-cbind(d,n)

zip2.2011<-subset(zip2,year==2011)
summary(zip2.2011)
zip2.2012<-subset(zip2,year==2012)
summary(zip2.2012)
zip2.2013<-subset(zip2,year==2013)
summary(zip2.2013)

zip2$price_in_state<-NULL
zip2$price_out_state<-NULL
?by
rsid.accessions <- tapply(zip$accessions, zip$rsid, sum)
zip2 <- zip
mat<-matrix(c("dist_to_nrs","median",
              "num_nrs_50mi","mean",
              "accessions","sum",
              "recruiters","sum",
              "recruiters_to_qma","median",
              "violent_crime","sum",
              "nonviolent_crime","sum",
              "qma","sum",
              "pop_2010","sum",
              "pop_dens","median",
              "mean_unemployment_rate","mean",
              "std_dv_unemployment","median",
              "avg_agi","median",
              "unemployment_comp","median",
              "per_returns_taxable","mean",
              "pensions_annuities_agi","mean",
              "agi_cat1","sum",
              "agi_cat2","sum",
              "agi_cat3","sum",
              "agi_cat4","sum",
              "agi_cat5","sum",
              "agi_cat6","sum",
              "six_fig","sum",
              "dist_to_uni","median",
              "num_uni_50mi","median",
              "total_schools","sum",
              "high_enrollment","max",
              "u_pop1","sum",
              "u_pop2","sum",
              "u_pop3","sum",
              "u_pop4","sum",
              "u_pop5","sum",
              "total_vet","sum",
```

```
            "total_nonvet","sum",
            "tot_m","sum",
            "tot_m_vet","sum",
            "tot_m_nonvet","sum",
            "m1","sum",
            "m1_v","sum",
            "m1_nv","sum",
            "m2","sum",
            "m2_v","sum",
            "m2_nv","sum",
            "m3","sum",
            "m3_v","sum",
            "m3_nv","sum",
            "m4","sum",
            "m4_v","sum",
            "m4_nv","sum",
            "m5","sum",
            "m5_v","sum",
            "m5_nv","sum",
            "tot_f","sum",
            "tot_f_vet","sum",
            "tot_f_nonvet","sum",
            "f1","sum",
            "f1_v","sum",
            "f1_nv","sum",
            "f2","sum",
            "f2_v","sum",
            "f2_nv","sum",
            "f3","sum",
            "f3_v","sum",
            "f3_nv","sum",
            "f4","sum",
            "f4_v","sum",
            "f4_nv","sum",
            "f5","sum",
            "f5_v","sum",
            "f5_nv","sum"), 70, 2, byrow=TRUE)

uni.2011<-subset(uni,uni$year==2011)
uni.2012<-subset(uni,uni$year==2012)
uni.2013<-subset(uni,uni$year==2013)

mat<-matrix(c("num_uni_50mi","median"), 1, 2, byrow=TRUE)
mout <- rsid.accessions
for (i in mat) { #1:nrow(mat)) {
   thing <- tapply (uni.2011[[mat[i,1]]], uni.2011$rsid, mat[i,2], na.rm=TRUE)
   cat ("thing", mat[i,1], "has length", length (thing), "\n")
   mout <- data.frame (mout, thing)
   names (mout)[ncol(mout)] <- mat[i,1]
}

thing11 <- tapply(uni.2011$num_uni_50mi, uni.2011$rsid, median, na.rm=TRUE)
thing11 <- data.frame(thing11)
thing12 <- tapply(uni.2012$num_uni_50mi, uni.2012$rsid, median, na.rm=TRUE)
thing12 <- data.frame(thing12)
```

```
thing13 <- tapply(uni.2013$num_uni_50mi, uni.2013$rsid, median, na.rm=TRUE)
thing13 <- data.frame(thing13)

write.table(thing11, "~/Desktop/NPS/Thesis/Data/Master Data Frame/Models/thing11.csv", sep=",")

###  2011 data
final.data11 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2011.csv")

###  2012 data
final.data12 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2012.csv")

###  2013 data
final.data13 <- read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/main_data_frame2013.csv")

write.table(all.rsid, "~/Desktop/NPS/Thesis/Data/Master Data Frame/all.rsid.csv", sep=",")
### Merge together all the zip code level data you just imported
#### zip<-rbind(final.data11, final.data12, final.data13)
summary(zip)
zip.2011<-subset(zip,year==2011)
summary(zip.2011)
zip.2012<-subset(zip,year==2012)
summary(zip.2012)
zip.2013<-subset(zip,year==2013)
summary(zip.2013)

zip<-rbind(zip.2011,zip.2012,zip.2013)
summary(zip)

write.table(zip, "~/Desktop/NPS/Thesis/Data/Master Data Frame/zip.csv", sep=",")
zip<-read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/zip.csv")

### USe this to justify what method I use to fill in missing values
hist((zip$violent_crime), ylim = c(0,2000), col="navyblue")  #Expontential
hist((zip$nonviolent_crime), ylim = c(0,1500), col="navyblue")  #Exponential
hist((zip$price_in_state), ylim = c(0,300), col="navyblue")  #Normal
hist((zip$price_out_state), ylim = c(0,300), col="navyblue") #normal
hist((zip$pop_dens), ylim = c(0,2000), col="navyblue")   #exponential
hist((zip$avg_agi), ylim=c(0,300), col="navyblue")  #exponential
hist(zip$per_returns_taxable, ylim=c(0,3000), xlim=c(.4,1.0), col="navyblue")  # normal
hist(zip$recruiters)  #Exponential
hist(zip$qma)  #Exponential
hist(zip$dist_to_uni)  #Exp
hist(zip$dist_to_nrs)#Exp
hist(zip$num_uni_50)#Exp
hist(zip$agi_cat1)#Exp
hist(zip$mean_unemployment_rate)  #normal
hist((zip$std_dv_unemployment), ylim=c(0,5000))  #exponential
hist((zip$unemployment_comp), ylim=c(0,3000))  #exponential
hist((zip$per_returns_taxable), ylim=c(0,5000), xlim=c(.4,1.0), col="navyblue")  #normal
hist((zip$pensions_annuities_agi), ylim=c(0,50), xlim=c(.4,1.0), col="navyblue") #exponential
hist((zip$num_uni_50mi), ylim=c(0,5000), col="navyblue") #exponential
hist(zip$recruiters_to_qma, 200)
```

```
### Replace 0s with NAs
zip$violent_crime[zip$violent_crime==0]<-NA
zip$nonviolent_crime[zip$nonviolent_crime==0]<-NA
length(zip$total_schools[zip$total_schools > 0])
98940-5262
94942-93678
1264/5262
rsid <- as.character(unique(zip$rsid))
### Recruiters  Substitution ###  6 NAs for recruiters = 6/98940
for (i in rsid) {
  zip.2011$recruiters[is.na (zip.2011$recruiters) & zip.2011$rsid == i] <-
    median (zip.2011$recruiters[!is.na (zip.2011$recruiters) & zip.2011$rsid == i])
}
for (i in rsid) {
  zip.2012$recruiters[is.na (zip.2012$recruiters) & zip.2012$rsid == i] <-
    median (zip.2012$recruiters[!is.na (zip.2012$recruiters) & zip.2012$rsid == i])
}
for (i in rsid) {
  zip.2013$recruiters[is.na (zip.2013$recruiters) & zip.2013$rsid == i] <-
    median (zip.2013$recruiters[!is.na (zip.2013$recruiters) & zip.2013$rsid == i])
}
### QMA Substitution ###  9 NAs for QMA = 9/98940 with 11221 zips with zero QMA = 11221/98940
for (i in rsid) {
  zip.2011$qma[is.na (zip.2011$qma) & zip.2011$rsid == i] <-
    median (zip.2011$qma[!is.na (zip.2011$qma) & zip.2011$rsid == i])
}
for (i in rsid) {
  zip.2012$qma[is.na (zip.2012$qma) & zip.2012$rsid == i] <-
    median (zip.2012$qma[!is.na (zip.2012$qma) & zip.2012$rsid == i])
}
for (i in rsid) {
  zip.2013$qma[is.na (zip.2013$qma) & zip.2013$rsid == i] <-
    median (zip.2013$qma[!is.na (zip.2013$qma) & zip.2013$rsid == i])
}

### Replace the PopDens NAs with medians  I have 9429 NAs across all years = 9429/98940
### 2011
for (i in rsid) {
  zip.2011$pop_dens[is.na (zip.2011$pop_dens) & zip.2011$rsid == i] <-
    median (zip.2011$pop_dens[!is.na (zip.2011$pop_dens) & zip.2011$rsid == i])
}
### 2012
for (i in rsid) {
  zip.2012$pop_dens[is.na (zip.2012$pop_dens) & zip.2012$rsid == i] <-
    median (zip.2012$pop_dens[!is.na (zip.2012$pop_dens) & zip.2012$rsid == i])
}
### 2013
for (i in rsid) {
  zip.2013$pop_dens[is.na (zip.2013$pop_dens) & zip.2013$rsid == i] <-
    median (zip.2013$pop_dens[!is.na (zip.2013$pop_dens) & zip.2013$rsid == i])
```

```
}


### It looks like we are missing about 10% of the data overall, so I will replace the NAs with the mean of
the
### data for that category by RSID, so I have 200 averages to pick from.  These are averages of zip codes
which make up a station

### avg_agi
for (i in rsid) {
 zip.2011$avg_agi[is.na (zip.2011$avg_agi) & zip.2011$rsid == i] <-
  median (zip.2011$avg_agi[!is.na (zip.2011$avg_agi) & zip.2011$rsid == i])
}
### unemployment comp
for (i in rsid) {
 zip.2011$unemployment_comp[is.na (zip.2011$unemployment_comp) & zip.2011$rsid == i] <-
  median (zip.2011$unemployment_comp[!is.na (zip.2011$unemployment_comp) & zip.2011$rsid == i])
}
### per returns taxable
for (i in rsid) {
 zip.2011$per_returns_taxable[is.na (zip.2011$per_returns_taxable) & zip.2011$rsid == i] <-
  mean (zip.2011$per_returns_taxable[!is.na (zip.2011$per_returns_taxable) & zip.2011$rsid == i])
}
### pensions annuities as part of agi
for (i in rsid) {
 zip.2011$pensions_annuities_agi[is.na (zip.2011$pensions_annuities_agi) & zip.2011$rsid == i] <-
  median (zip.2011$pensions_annuities_agi[!is.na (zip.2011$pensions_annuities_agi) & zip.2011$rsid ==
i])
}
### cat1
for (i in rsid) {
 zip.2011$agi_cat1[is.na (zip.2011$agi_cat1) & zip.2011$rsid == i] <-
  median (zip.2011$agi_cat1[!is.na (zip.2011$agi_cat1) & zip.2011$rsid == i])
}
### cat2
for (i in rsid) {
 zip.2011$agi_cat2[is.na (zip.2011$agi_cat2) & zip.2011$rsid == i] <-
  median (zip.2011$agi_cat2[!is.na (zip.2011$agi_cat2) & zip.2011$rsid == i])
}
### cat3
for (i in rsid) {
 zip.2011$agi_cat3[is.na (zip.2011$agi_cat3) & zip.2011$rsid == i] <-
  median (zip.2011$agi_cat3[!is.na (zip.2011$agi_cat3) & zip.2011$rsid == i])
}
### cat4
for (i in rsid) {
 zip.2011$agi_cat4[is.na (zip.2011$agi_cat4) & zip.2011$rsid == i] <-
  median (zip.2011$agi_cat4[!is.na (zip.2011$agi_cat4) & zip.2011$rsid == i])
}
### cat5
for (i in rsid) {
```

```
 zip.2011$agi_cat5[is.na (zip.2011$agi_cat5) & zip.2011$rsid == i] <-
   median (zip.2011$agi_cat5[!is.na (zip.2011$agi_cat5) & zip.2011$rsid == i])
}
### cat6
for (i in rsid) {
 zip.2011$agi_cat6[is.na (zip.2011$agi_cat6) & zip.2011$rsid == i] <-
   median (zip.2011$agi_cat6[!is.na (zip.2011$agi_cat6) & zip.2011$rsid == i])
}
### 2012 ###
### avg_agi
for (i in rsid) {
 zip.2012$avg_agi[is.na (zip.2012$avg_agi) & zip.2012$rsid == i] <-
   median (zip.2012$avg_agi[!is.na (zip.2012$avg_agi) & zip.2012$rsid == i])
}
### unemployment comp
for (i in rsid) {
 zip.2012$unemployment_comp[is.na (zip.2012$unemployment_comp) & zip.2012$rsid == i] <-
   median (zip.2012$unemployment_comp[!is.na (zip.2012$unemployment_comp) & zip.2012$rsid == i])
}
### per returns taxable
for (i in rsid) {
 zip.2012$per_returns_taxable[is.na (zip.2012$per_returns_taxable) & zip.2012$rsid == i] <-
   mean (zip.2012$per_returns_taxable[!is.na (zip.2012$per_returns_taxable) & zip.2012$rsid == i])
}
### pensions annuities as part of agi
for (i in rsid) {
 zip.2012$pensions_annuities_agi[is.na (zip.2012$pensions_annuities_agi) & zip.2012$rsid == i] <-
   median (zip.2012$pensions_annuities_agi[!is.na (zip.2012$pensions_annuities_agi) & zip.2012$rsid ==
i])
}
### cat1
for (i in rsid) {
 zip.2012$agi_cat1[is.na (zip.2012$agi_cat1) & zip.2012$rsid == i] <-
   median (zip.2012$agi_cat1[!is.na (zip.2012$agi_cat1) & zip.2012$rsid == i])
}
### cat2
for (i in rsid) {
 zip.2012$agi_cat2[is.na (zip.2012$agi_cat2) & zip.2012$rsid == i] <-
   median (zip.2012$agi_cat2[!is.na (zip.2012$agi_cat2) & zip.2012$rsid == i])
}
### cat3
for (i in rsid) {
 zip.2012$agi_cat3[is.na (zip.2012$agi_cat3) & zip.2012$rsid == i] <-
   median (zip.2012$agi_cat3[!is.na (zip.2012$agi_cat3) & zip.2012$rsid == i])
}
### cat4
for (i in rsid) {
 zip.2012$agi_cat4[is.na (zip.2012$agi_cat4) & zip.2012$rsid == i] <-
   median (zip.2012$agi_cat4[!is.na (zip.2012$agi_cat4) & zip.2012$rsid == i])
}
### cat5
```

```
for (i in rsid) {
 zip.2012$agi_cat5[is.na (zip.2012$agi_cat5) & zip.2012$rsid == i] <-
  median (zip.2012$agi_cat5[!is.na (zip.2012$agi_cat5) & zip.2012$rsid == i])
}
### cat6
for (i in rsid) {
 zip.2012$agi_cat6[is.na (zip.2012$agi_cat6) & zip.2012$rsid == i] <-
  median (zip.2012$agi_cat6[!is.na (zip.2012$agi_cat6) & zip.2012$rsid == i])
}

### Crime All Years
for (i in rsid) {
 zip.2011$violent_crime[is.na (zip.2011$violent_crime) & zip.2011$rsid == i] <-
  median (zip.2011$violent_crime[!is.na (zip.2011$violent_crime) & zip.2011$rsid == i])
}

for (i in rsid) {
 zip.2011$nonviolent_crime[is.na (zip.2011$nonviolent_crime) & zip.2011$rsid == i] <-
  median (zip.2011$nonviolent_crime[!is.na (zip.2011$nonviolent_crime) & zip.2011$rsid == i])
}
for (i in rsid) {
 zip.2012$violent_crime[is.na (zip.2012$violent_crime) & zip.2012$rsid == i] <-
  median (zip.2012$violent_crime[!is.na (zip.2012$violent_crime) & zip.2012$rsid == i])
}

for (i in rsid) {
 zip.2012$nonviolent_crime[is.na (zip.2012$nonviolent_crime) & zip.2012$rsid == i] <-
  median (zip.2012$nonviolent_crime[!is.na (zip.2012$nonviolent_crime) & zip.2012$rsid == i])
}
for (i in rsid) {
 zip.2013$violent_crime[is.na (zip.2013$violent_crime) & zip.2013$rsid == i] <-
  median (zip.2013$violent_crime[!is.na (zip.2013$violent_crime) & zip.2013$rsid == i])
}

for (i in rsid) {
 zip.2013$nonviolent_crime[is.na (zip.2013$nonviolent_crime) & zip.2013$rsid == i] <-
  median (zip.2013$nonviolent_crime[!is.na (zip.2013$nonviolent_crime) & zip.2013$rsid == i])
}

### Unemployment Rate and Std Dv Unemployment Rate
for (i in rsid) {
 zip.2011$mean_unemployment_rate[is.na (zip.2011$mean_unemployment_rate) & zip.2011$rsid == i]
<-
  median (zip.2011$mean_unemployment_rate[!is.na (zip.2011$mean_unemployment_rate) &
zip.2011$rsid == i])
}
for (i in rsid) {
 zip.2012$mean_unemployment_rate[is.na (zip.2012$mean_unemployment_rate) & zip.2012$rsid == i]
<-
  median (zip.2012$mean_unemployment_rate[!is.na (zip.2012$mean_unemployment_rate) &
zip.2012$rsid == i])
```

```
}
for (i in rsid) {
 zip.2013$mean_unemployment_rate[is.na (zip.2013$mean_unemployment_rate) & zip.2013$rsid == i]
<-
   median (zip.2013$mean_unemployment_rate[!is.na (zip.2013$mean_unemployment_rate) &
zip.2013$rsid == i])
}
for (i in rsid) {
 zip.2011$std_dv_unemployment[is.na (zip.2011$std_dv_unemployment) & zip.2011$rsid == i] <-
   median (zip.2011$std_dv_unemployment[!is.na (zip.2011$std_dv_unemployment) & zip.2011$rsid ==
i])
}
for (i in rsid) {
 zip.2012$std_dv_unemployment[is.na (zip.2012$std_dv_unemployment) & zip.2012$rsid == i] <-
   median (zip.2012$std_dv_unemployment[!is.na (zip.2012$std_dv_unemployment) & zip.2012$rsid ==
i])
}
for (i in rsid) {
 zip.2013$std_dv_unemployment[is.na (zip.2013$std_dv_unemployment) & zip.2013$rsid == i] <-
   median (zip.2013$std_dv_unemployment[!is.na (zip.2013$std_dv_unemployment) & zip.2013$rsid ==
i])
}

### Write a piece of code that fills in a 0 for each NA block under price_in_state and price_out_state
### where the corresponding total_schools zip code is greater than 0
zip$price_in_state[zip$price_in_state==0]<-NA
zip$price_out_state[zip$price_out_state==0]<-NA

zip$price_in_state[is.na (zip$price_in_state) & zip$total_schools == 0] <- 99
zip$price_out_state[is.na (zip$price_out_state) & zip$total_schools == 0] <- 99

zip$price_in_state[is.na (zip$price_in_state) & zip$total_schools > 0] <- 0
zip$price_out_state[is.na (zip$price_out_state) & zip$total_schools > 0] <- 0

### Then write a piece of code that takes those zeros and replaces them with the mean of all the zip
codes tuitions
### in that RSID
length(zip$price_in_state[zip$price_in_state==99])
length(zip$price_out_state[zip$price_out_state==99])
length(zip$price_in_state[is.na(zip$price_in_state)])
length(zip$price_out_state[is.na(zip$price_out_state)])
for (i in rsid) {
 zip.2011$price_in_state[is.na (zip.2011$price_in_state) & zip.2011$rsid == i] <-
   mean(zip.2011$price_in_state[!is.na (zip.2011$price_in_state) & zip.2011$total_schools > 0 &
zip.2011$rsid == i])
}

for (i in rsid) {
 zip.2011$price_out_state[is.na (zip.2011$price_out_state) & zip.2011$rsid == i] <-
   mean(zip.2011$price_out_state[!is.na (zip.2011$price_out_state) & zip.2011$total_schools > 0 &
zip.2011$rsid == i])
```

```
}
for (i in rsid) {
  zip.2012$price_in_state[is.na (zip.2012$price_in_state) & zip.2012$rsid == i] <-
    mean(zip.2012$price_in_state[!is.na (zip.2012$price_in_state) & zip.2012$total_schools > 0 &
zip.2012$rsid == i])
}

for (i in rsid) {
  zip.2012$price_out_state[is.na (zip.2012$price_out_state) & zip.2012$rsid == i] <-
    mean(zip.2012$price_out_state[!is.na (zip.2012$price_out_state) & zip.2012$total_schools > 0 &
zip.2012$rsid == i])
}
for (i in rsid) {
  zip.2013$price_in_state[is.na (zip.2013$price_in_state) & zip.2013$rsid == i] <-
    mean(zip.2013$price_in_state[!is.na (zip.2013$price_in_state) & zip.2013$total_schools > 0 &
zip.2013$rsid == i])
}

for (i in rsid) {
  zip.2013$price_out_state[is.na (zip.2013$price_out_state) & zip.2013$rsid == i] <-
    mean(zip.2013$price_out_state[!is.na (zip.2013$price_out_state) & zip.2013$total_schools > 0 &
zip.2013$rsid == i])
}

### Group by NRD and fill in missing values for the price in state (300), price out state (300), violent crime
(709), and nonviolent crime (39)
nrd <- as.character(unique(zip$nrd))
for (i in nrd) {
  zip.2011$price_in_state[is.na (zip.2011$price_in_state) & zip.2011$nrd == i] <-
    mean(zip.2011$price_in_state[!is.na (zip.2011$price_in_state) & zip.2011$total_schools > 0 &
zip.2011$nrd == i])
}

for (i in nrd) {
  zip.2011$price_out_state[is.na (zip.2011$price_out_state) & zip.2011$nrd == i] <-
    mean(zip.2011$price_out_state[!is.na (zip.2011$price_out_state) & zip.2011$total_schools > 0 &
zip.2011$nrd == i])
}
for (i in nrd) {
  zip.2012$price_in_state[is.na (zip.2012$price_in_state) & zip.2012$nrd == i] <-
    mean(zip.2012$price_in_state[!is.na (zip.2012$price_in_state) & zip.2012$total_schools > 0 &
zip.2012$nrd == i])
}

for (i in nrd) {
  zip.2012$price_out_state[is.na (zip.2012$price_out_state) & zip.2012$nrd == i] <-
    mean(zip.2012$price_out_state[!is.na (zip.2012$price_out_state) & zip.2012$total_schools > 0 &
zip.2012$nrd == i])
}
for (i in nrd) {
  zip.2013$price_in_state[is.na (zip.2013$price_in_state) & zip.2013$nrd == i] <-
```

```
   mean(zip.2013$price_in_state[!is.na (zip.2013$price_in_state) & zip.2013$total_schools > 0 &
zip.2013$nrd == i])
}

for (i in nrd) {
  zip.2013$price_out_state[is.na (zip.2013$price_out_state) & zip.2013$nrd == i] <-
    mean(zip.2013$price_out_state[!is.na (zip.2013$price_out_state) & zip.2013$total_schools > 0 &
zip.2013$nrd == i])
}

#### Good, it got rid of the remaining NAs, now onto crime.

### Crime All Years
for (i in nrd) {
  zip.2011$violent_crime[is.na (zip.2011$violent_crime) & zip.2011$nrd == i] <-
    median (zip.2011$violent_crime[!is.na (zip.2011$violent_crime) & zip.2011$nrd == i])
}

for (i in nrd) {
  zip.2011$nonviolent_crime[is.na (zip.2011$nonviolent_crime) & zip.2011$nrd == i] <-
    median (zip.2011$nonviolent_crime[!is.na (zip.2011$nonviolent_crime) & zip.2011$nrd == i])
}
for (i in nrd) {
  zip.2012$violent_crime[is.na (zip.2012$violent_crime) & zip.2012$nrd == i] <-
    median (zip.2012$violent_crime[!is.na (zip.2012$violent_crime) & zip.2012$nrd == i])
}

for (i in nrd) {
  zip.2012$nonviolent_crime[is.na (zip.2012$nonviolent_crime) & zip.2012$nrd == i] <-
    median (zip.2012$nonviolent_crime[!is.na (zip.2012$nonviolent_crime) & zip.2012$nrd == i])
}
for (i in nrd) {
  zip.2013$violent_crime[is.na (zip.2013$violent_crime) & zip.2013$nrd == i] <-
    median (zip.2013$violent_crime[!is.na (zip.2013$violent_crime) & zip.2013$nrd == i])
}

for (i in nrd) {
  zip.2013$nonviolent_crime[is.na (zip.2013$nonviolent_crime) & zip.2013$nrd == i] <-
    median (zip.2013$nonviolent_crime[!is.na (zip.2013$nonviolent_crime) & zip.2013$nrd == i])
}

### Alright, you have no more NAs, save for the 2013 financial data

####  Take the zip file and break it into a test set and training set
east.zip<-subset(zip,region=="EAST")
west.zip<-subset(zip,region=="WEST")
national.test<-subset(zip,year==2013)
east.test<-subset(east.zip,east.zip$year==2013)
west.test<-subset(west.zip,west.zip$year==2013)

national.train<-subset(zip,year<2013)
```

```
east.train<-subset(east.zip,east.zip$year<2013)
west.train<-subset(west.zip,west.zip$year<2013)

write.table(national.train, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/ZIP/national.train.csv", sep=",")
write.table(east.train, "~/Desktop/NPS/Thesis/Data/Master Data Frame/Models/ZIP/east.train.csv",
sep=",")
write.table(west.train, "~/Desktop/NPS/Thesis/Data/Master Data Frame/Models/ZIP/west.train.csv",
sep=",")

write.table(national.test, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/ZIP/Test/national.test.csv", sep=",")
write.table(east.test, "~/Desktop/NPS/Thesis/Data/Master Data Frame/Models/ZIP/Test/east.test.csv",
sep=",")
write.table(west.test, "~/Desktop/NPS/Thesis/Data/Master Data Frame/Models/ZIP/Test/west.test.csv",
sep=",")

all.rsid<-read.csv("~/Desktop/NPS/Thesis/Data/Master Data Frame/all.rsid.csv")
east.rsid<-subset(all.rsid,region=="EAST")
west.rsid<-subset(all.rsid,region=="WEST")
national.rsid.test<-subset(all.rsid,year==2013)
east.rsid.test<-subset(east.rsid,east.rsid$year==2013)
west.rsid.test<-subset(west.rsid,west.rsid$year==2013)

national.rsid.train<-subset(all.rsid,year<2013)
east.rsid.train<-subset(east.rsid,east.rsid$year<2013)
west.rsid.train<-subset(west.rsid,west.rsid$year<2013)

write.table(national.rsid.train, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/national.rsid.train.csv", sep=",")
write.table(east.rsid.train, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/east.rsid.train.csv", sep=",")
write.table(west.rsid.train, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/west.rsid.train.csv", sep=",")

write.table(national.rsid.test, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/Test/national.rsid.test.csv", sep=",")
write.table(east.rsid.test, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/Test/east.rsid.test.csv", sep=",")
write.table(west.rsid.test, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/GLM/Test/west.rsid.test.csv", sep=",")

### Now that you have data for national level, East, West models with zip level data and national level,
East, and West models
### with RSID level data, you need to bin the responses by quartile and save that data set for model
building.  You will have
### to run quartiles 3 times, because they may be different based on distributions.
summary(all.rsid$accessions)
summary(east.rsid$accessions)
summary(west.rsid$accessions)
### Do it at the National Level
```

```
all.rsid.d<-all.rsid
decLocations <- quantile(all.rsid.d$accessions, probs = seq(0.1,0.9,by=0.1))
dec<-findInterval(all.rsid.d$accessions,c(0,decLocations, Inf))
all.rsid.d$accessions<-dec

east.rsid.d<-east.rsid
### East Region
decLocations <- quantile(east.rsid.d$accessions, probs = seq(0.1,0.9,by=0.1))
dec<-findInterval(east.rsid.d$accessions,c(0,decLocations, Inf))
east.rsid.d$accessions<-dec

west.rsid.d<-west.rsid
### West Region
decLocations <- quantile(west.rsid.d$accessions, probs = seq(0.1,0.9,by=0.1))
dec<-findInterval(west.rsid.d$accessions,c(0,decLocations, Inf))
west.rsid.d$accessions<-dec

national.rsid.dtest<-subset(all.rsid.d,year==2013)
east.rsid.dtest<-subset(east.rsid.d,east.rsid.d$year==2013)
west.rsid.dtest<-subset(west.rsid.d,west.rsid.d$year==2013)

national.rsid.dtrain<-subset(all.rsid.d,year<2013)
east.rsid.dtrain<-subset(east.rsid.d,east.rsid.d$year<2013)
west.rsid.dtrain<-subset(west.rsid.d,west.rsid.d$year<2013)

### Now find the quartiles
### Do it at the National Level
all.rsid.q<-all.rsid
decLocations <- quantile(all.rsid.q$accessions, probs = seq(0.25,.75,by=.25))
dec<-findInterval(all.rsid.q$accessions,c(0,decLocations, Inf))
all.rsid.q$accessions<-dec

east.rsid.q<-east.rsid
### East Region
decLocations <- quantile(east.rsid.q$accessions, probs = seq(0.25,0.75,by=0.25))
dec<-findInterval(east.rsid.q$accessions,c(0,decLocations, Inf))
east.rsid.q$accessions<-dec

west.rsid.q<-west.rsid
### West Region
decLocations <- quantile(west.rsid.q$accessions, probs = seq(0.25,0.75,by=0.25))
dec<-findInterval(west.rsid.q$accessions,c(0,decLocations, Inf))
west.rsid.q$accessions<-dec

national.rsid.qtest<-subset(all.rsid.q,year==2013)
east.rsid.qtest<-subset(east.rsid.q,east.rsid.q$year==2013)
west.rsid.qtest<-subset(west.rsid.q,west.rsid.q$year==2013)

national.rsid.qtrain<-subset(all.rsid.q,year<2013)
east.rsid.qtrain<-subset(east.rsid.q,east.rsid.q$year<2013)
west.rsid.qtrain<-subset(west.rsid.q,west.rsid.q$year<2013)
```

```
### Now, write it all to csv
### Decile data
write.table(national.rsid.dtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/national.rsid.dtrain.csv", sep=",")
write.table(east.rsid.dtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/east.rsid.dtrain.csv", sep=",")
write.table(west.rsid.dtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/west.rsid.dtrain.csv", sep=",")

write.table(national.rsid.dtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/national.rsid.dtest.csv", sep=",")
write.table(east.rsid.dtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/east.rsid.dtest.csv", sep=",")
write.table(west.rsid.dtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/west.rsid.dtest.csv", sep=",")

### Quartile Data
write.table(national.rsid.qtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/national.rsid.qtrain.csv", sep=",")
write.table(east.rsid.qtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/east.rsid.qtrain.csv", sep=",")
write.table(west.rsid.qtrain, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/west.rsid.qtrain.csv", sep=",")

write.table(national.rsid.qtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/national.rsid.qtest.csv", sep=",")
write.table(east.rsid.qtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/east.rsid.qtest.csv", sep=",")
write.table(west.rsid.qtest, "~/Desktop/NPS/Thesis/Data/Master Data
Frame/Models/MLR/Test/west.rsid.qtest.csv", sep=",")

temps <- read.csv("~/Desktop/NPS/8th QTR/Manpower Models/temps.csv")
as.vector(temps)
temps$time <- NULL
names(temps)
sma<-tapply(temps$temp, temps$day, mean)
plot(sma)
write.table(sma, "~/Desktop/NPS/8th QTR/Manpower Models/sma.csv", sep=",")

y<-tapply(final.data11$accessions, final.data11$rsid, sum)
typeof(temps)
mp.lm <- lm(cost~depth, data = mp)
summary(mp.lm)
par(mfrow=c(2,2))
plot(mp.lm, add.smooth=F)
shapiro.test(residuals(mp.lm))
plot(fitted(mp.lm), residuals(mp.lm), xlab = "Fitted", ylab = 'Residuals', main = 'West Region 2 Yr Model
Residual v Fitted Values')
abline(h = 0,col ='red')
```

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF REFERENCES


Athnet. (2015). Division I universities. Retrieved February 2015 from http://www.athleticscholarships.net/division-1-colleges-schools.htm

Chin, D.G. (2015). *Innovative statistical methods for public health.* New York, NY: Springer.

Commander Navy Recruiting Command (CNRC). (2011, May 17). *Navy recruiting manual, recruiting operatons*. Millington TN: CNRC. Retrieved from COMNAVCRUITCOMINST 1130.8J: http://www.cnrc.navy.mil/Publications/Directives/1130.8/1130.8J_VOL%20I_Recruiting%20Operations-CH8.pdf

Cook, L. (2014, September 22). Is the college admissions bubble about to burst? *U.S. News*. Retrieved from http://www.usnews.com/news/blogs/data-mine/2014/09/22/is-the-college-admissions-bubble-about-to-burst

Dynarski, S. (2000). Hope for whom? Financial aid for the middle class and its impact on college attendance. *NBER Working Papers, 7756*, 4.

Faraway, J. (2006). *Extending the linear model with R.* Boca Raton, FL: Taylor and Francis Group.

Federal Bureau of Investigation. (2015). Uniform crime reports. Retrieved from https://www.fbi.gov/about-us/cjis/ucr/ucr

Feeney, N. (2014, June 29). Pentagon: 7 in10 youths would fail to qualify for military service, *Time*. Retrieved from http://time.com/2938158/youth-fail-to-qualify-military-service/

Flynn, M. (2009, December). More flexible GLMs, zero-inflated models and hybrid models. *Casualty Actuarial Society E-Forum*, 149.

Gibson, J. (2009). *Zip Code valuation study technical report: Predicting Army accessions.* Arlington, VA: DOD.

Gill, J. (2004). What to do when your hessian is not invertible. *Sociological Methods and Research, 32*(4), 2.

Gilmour, S. G. (1996). The interpretation of Mallow's Cp-statistic. *Journal of the Royal Statistical Society , 45*(1), 49-56.

Internal Revenue Service. (2015). *Internal Revenue Service SOI*. Retrieved from http://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-ZIP-Code-Data-(SOI)

Ismail, N. (2013, April). Estimation of claim count data using negative binomial. *Casual Actuarial Society*. Retrieved from https://www.casact.org/pubs/forum/13spforum/Ismail%20Zamani.pdf

Jackman, S. (2015). Pscl: Classes and methods for R developed in the political science computational laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.4.9. URL http://pscl.stanford.edu/

Jackson, S. (2015). *Utilizing socio-economic factors to evaluate recruiting potential for a US Army recruiting company.* Austin, TX: University of Texas at Austin.

Janowitz, M. (1973). The social demography of the all-volunteer armed force. *The Annals of the American Academy of Political and Social Science, 406*(1), 87.

Kleykamp, M. A. (2006). College, jobs, or the military? Enlistment during a time of war. *Social Science Quarterly, 87*(2), 273.

Lee, J. M. (2007). Trends in hospitalizations for diabetes among children and young adults in the United States. *Diabetes Care, 30*(12), 35.

Lumley, T. using Fortran code by Alan Miller (2009). Leaps: Regression subset selection. R package version 2.9. Retrieved from http://CRAN.R-project.org/package=leaps

Malone, L. (2009). *Review of methods for the district allocation of prior service recruiters.* Arlington, VA: Center for Naval Analysis.

Marmion, W. (2015). *Evaluating and improving the SAMA (Segmentation Analysis and Market Assessment) recruiting model* (Master's thesis). Retrieved from Calhoun http://calhoun.nps.edu/bitstream/handle/10945/45894.

Marsh, P. (2011). *Department of defense youth poll wave 20.* Alexandria, VA: Defense Human Resources Activity.

NCAA. (2015). *NCAA Divsion I universities*. Retrieved September 15, 2015, from NCAA Divsion I Universities: http://www.ncaa.org/about?division=d1

*New York Times*. (2012, January 14). What percent are you? Retrieved: http://www.nytimes.com/interactive/2012/01/15/business/one-percent-map.html?_r=0

Pinelis, Y. K., Schmitz, E., Miller, Z. & Rebhan, E. (2011). *An analysis of Navy recruiting goal allocation models.* Arlington, VA: CNA.

Piza, E. (2012). Using Poisson and negative binomial regression models to measure the influence of risk on crime incident counts. *Risk Terrain Modeling*.

Poole, M. (1970, July 10). *The assumptions of the linear regression model.* Retrieved from http://people.uleth.ca/~towni0/PooleOfarrell71.pdf

Princeton University. (2015). *Data and Statistical Services*. Retrieved from http://dss.princeton.edu/online_help/analysis/regression_intro.htm

R Development Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: The R Foundation for Statistical Computing.

Rodriguez, G. (2013, November 6). Models for count data with overdispersion. Retrieved from Princeton University http://data.princeton.edu/wws509/notes/c4a.pdf

Rossum, G. (1995, May). PYTHON Tutorial. *Technical Report CS-R9526.* Amsterdam: CWI TR.

Schacherer, R. A. (2005). The conditions affecting military enlistments. *The Public Purpose, III*, 76-82.

United States Census Bureau. (2015). *United States Census Bureau 3-year American community survey*. Retrieved March 2015, from United States Census Bureau 3 year American Community Survey: https://www.census.gov/programs-surveys/acs/

United States Census. (2010, January). *Data.gov*. Retrieved February 2015, from Data.gov: http://www.data.gov

United States Department of Labor. (2014, June 12). *Bureau of Labor Statistics*. Retrieved August 17, 2015, from Bureau of Labor Statistics: http://www.bls.gov/cps/cps_htgm.htm#unemployed

Venables, W. N. & Ripley, B. D. (2002) Modern applied statistics with S. Fourth Edition. New York: Springer.

Williams, T. (2014). *Understanding factors influencing Navy recruiting production.* Monterey, CA: Naval Postgraduate School.

Wilson, M. J. (1999). *YATS 1999 Propensity and advertising report.* Arlington, VA: Defense Manpower Data Center.

Woods & Poole. Long-term county forecasts of employment, population, income, retail sales & households. *Woods & Poole Economics, Inc. Long-term county forecasts of employment, population, income, retail sales & households.* 2013. Retrieved February 3, 2015, from Woods & Poole: http://www.woodsandpoole.com/

Zeileis, A. (2008). Regression models for count data in R. *Journal of Statistical Softwar, 27* (8), 7.

Zhu Wang, with contributions from Achim Zeileis, Simon Jackman, Brian Ripley, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Gil Chu and Patrick Breheny (2015). mpath: Regularized linear models. R package version 0.1-19. http://CRAN.R-project.org/package=mpath

ZIP Boundary. (2014). *ZIP boundary FAQs*. Retrieved from http://www.zipboundary.com/zipcode_faqs.html

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California